Anatomy of missense variants in health and disease towards better impact prediction with a focus on titinopathies

Laddach, Anna Christine

Awarding institution:
King's College London

Anatomy of Missense Variants in Health and Disease

Towards Better Impact Prediction with a Focus on Titinopathies

Anna Christine Laddach

Randall Centre for Cell & Molecular Biophysics

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

April 2019

Acknowledgements

The first thank you must go to my supervisor Prof. Franca Fraternali, not only for supporting me throughout this PhD, but also for all the wonderful opportunities she has given me. No other supervisor could be as colourful (literally) or have as much enthusiasm; nor would the lab be stocked with the essential fuel, Nespresso.

Secondly, I thank past and present Fraternali group members; Joseph, Irene, Marius, Jamie, Christian, Rasha, Molly and Sun. It has been a privilege to work with such genuine people. Special acknowledgements must go to Joseph, for fascinating bioinformatics-related discussion, and Irene, for teaching me how to run MD simulations. You are both amazing! Also, life in the lab would not have been the same without Jamie's electric car stories...

Thanks also go to my second supervisor, Mathias Gautel, and the Gautel lab; in particular to Roksana and Martin for providing biophysical data and crystal structures for titin domains. This work, especially Chapter 4, would not have been possible without you; I am sure the wet lab would have blown up, had I stepped inside it!

To my wonderful parents, Liz and Alfons who went through the arduous task of proofreading this whole thesis. I am incredibly grateful! Here, I must also mention those Tesco vouchers you gave me which enabled me to do my first OU course (yes, Tesco Clubcard points really have played a role in my education). Also, thank you to my sister Katrina; what would I do without those delicious chocolate truffles?

Finally, the biggest acknowledgement must go to my husband Roman, who has been with me throughout my journey from music to science. Thank you for being an optimist, for supporting me, for being the quirkiest, most curious human being I know, and for making me coffee! Kocham cię.

As a parting note, I feel this journey has shown me that one is not at the mercy of fate (although to understand this I would probably have to study quantum mechanics), one can make drastic changes, and yes, it is not necessarily easy, but there are people who will help you in this adventure.

> “We have to build the Republic of Heaven where we are, because for us, there is no elsewhere.”
>
> *-Philip Pullman, The Amber Spyglass*

Abstract

The genomic revolution has brought about large advances in the identification of disease-associated variants. However, despite the recent explosion of genetic data, the problem of missing heritability persists. Variants with low penetrance remain difficult to identify, as do variants which are rare or unique to a single individual. To fully understand disease mechanisms and design targeted therapies, the molecular mechanisms underlying the pathogenic effects of such variants must be clarified. A prime example of missense variants which are difficult to classify is provided by those which localise the Titin gene, a number of which are associated with titinopathies. Due to titin's large size, even the majority of healthy individuals possess one or more rare titin missense variants. This results in the paradox that rare titin variants are commonly found; therefore, pathogenicity cannot be inferred from frequency alone. To address this issue we have created a web application, TITINdb (http://fraternalilab.kcl.ac.uk/TITINdb/), which integrates structural, variant, sequence and isoform information along with precomputed in-silico analyses, in order to facilitate the prioritisation of variants for further wet-lab investigation.

Recently available databases allow access to missense variant data on an unprecedented scale. We sought to harness this information to better understand the characteristics of variants associated with health and disease, through a large scale-analysis of population variants from the gnomAD database, as well as disease-associated variants (ClinVar) and somatic cancer-associated variants (COSMIC). Here we established that variants from each data set target distinct functional pathways and proteomics features. In order to accomplish this analysis, we created a database, web interface and REST API, ZoomVar (http://fraternalilab.kcl.ac.uk/ZoomVar/), to allow for the mapping of variants to a 3D integrated protein-protein interaction network and calculation of the regional enrichment of missense variants.

Despite the multitude of features which are able to segregate deleterious from neutral missense variants, a number of problem cases remain. This motivated us to investigate whether features extracted from molecular dynamics simulations could improve predictions of variant deleteriousness. To accomplish this we constructed a dataset of rare population and deleterious titin variants, and created machine-learning (random forest) based models of variant impact. We show that dynamics-based features are able to segregate the majority of disease-associated titin variants from population variants. Ultimately, we believe a collaborative framework for the sharing of mutant and wild-type trajectories must be set up; both to enable investigation into the possible benefits of using dynamics-based features, and to harness their power.

Published Articles

Publications which form part of this work:

Laddach, A., Gautel, M., and Fraternali, F. (2017). TITINdb-a computational tool to assess titin's role as a disease gene. *Bioinformatics (Oxford, England)*, 33(21):3482–3485.
This paper is presented in Chapter 2

Laddach, A., Ng, J. C.-F., Chung, S. S., and Fraternali, F. (2018). Genetic variants and protein-protein interactions: a multidimensional network-centric view. *Current opinion in structural biology*, 50:82–90.
Chapter 1 contains material from this paper

Additional publications:

Laddach, A., Chung, S. S., and Fraternali, F. (2018). Prediction of Protein-Protein Interactions: Looking Through the Kaleidoscope. In: Ranganathan, S., Nakai, K., Schönbach C. and Gribskov, M. (eds.) *Encyclopedia of Bioinformatics and Computational Biology*, vol. 2, pp. 834 - 848. Oxford: Elsevier.

Chung, S. S., Laddach, A., Thomas, N. S. B., and Fraternali, F. (2018). Short loop motif profiling of protein interaction networks in acute myeloid leukaemia. *bioRxiv*.

Vigilante, A., Laddach, A., Moens, M., Meleckyte, R., Leha, A., Ghahramani, A., Culley, O. J., Kathuria, A., Hurling, C., Vickers, A., Tewary, M., Zandstra, P., HipSci Consortium, Durbin, R., Fraternali, F., Stegle, O., Birney, E., Luscombe, N. M., Danovi, D., and Watt, F. M. (2019). Identifying extrinsic versus intrinsic drivers of variation in cell behaviour in human iPS cell lines from healthy donors. *Cell reports*, 26(8):2078—87.

Table of contents

List of figures

List of tables

Chapter 1

Introduction and background to the work

The basis of life is code. In this manner, all living beings are not so dissimilar to computer programs. However, unlike such programs, at least those which are reasonably well documented, living beings do not come with an instruction manual. Rather than binary code, as in computing, four different base pairs are possible. Errors in this code give rise to phenotypic variety. This is a double-edged sword, as such variety can result in both selective advantages and lend a species the flexibility to adapt to environmental changes, however it can also lead to detrimental variants. A number of different types of variants can occur, parts of the code can be deleted or extra pieces of code can be inserted, or reading of the code can be brought prematurely to a halt. These types of variants are called indels (insertions and deletions) and truncating variants. Here we focus instead on single nucleotide variants; those variants where one nucleotide is swapped for another one. More specifically we focus on non-synonymous single nucleotide variants (nsSNVs). These are variants which occur in protein coding regions of the genome and result in a change in the amino acid sequence of a protein. Such variants can also be termed single amino acid variants (SAVs), and are commonly notated as a string in which the letter representing the wild-type amino acid is followed by the position within the protein or domain and finally the mutant amino acid (e.g. T560M).

As no instruction manual exists we must endeavour to decipher the code, and, importantly, to understand which variants lead to disease and why. Huge progress has been made since the first

human genome sequence was solved in 2004 (International Human Genome Sequencing Consortium, 2004), an endeavour which took 10 years to complete. Of particular importance was the development of next-generation sequencing (NGS) technology. This has massively decreased the cost and time for sequencing. Now genetic data exists for a huge number of individuals - over 100,000 in the gnomAD database (Lek et al., 2016), and the bottleneck has shifted to analysis and interpretation of this data. Initially, research focussed on so-called low hanging fruits and identified variants with easily detectable disease associations. However, for a large number of diseases and phenotypes thought to be heritable, the genetic component remains unidentified, resulting in the so-termed problem of "missing heritability" (Manolio et al., 2009). With the enormous amount of data which is now available, it becomes imperative to go beyond identifying the most obvious disease-associated variants. In this light, it is essential to understand how disease-associated missense variants differ from neutral variants in their localisation to protein structure and impact on protein function. Moreover, this knowledge can be used to inform prediction.

This work endeavours to reach an improved understanding of the properties of human missense variants in health and disease, and to contribute to the development of computational variant impact predictors. Although at times we take a proteome-wide approach, the majority of this work is focussed on the protein titin, so named for its titanic size (the longest isoform is 35,991 amino acids in length). Variants which localise to this protein are particularly difficult to classify, although several disease-associations have already been uncovered. As a large number of titin variants of unknown significance are found, improved impact prediction is particularly essential for their assessment. Moreover, we believe that the methods used here offer scope for further development and can be applied to other proteins with problem-case variants, and ultimately contribute to the elucidation of missing heritability.

1.1 Detecting phenotype-associated nsSNVs

Understanding the impact of nsSNVs on phenotype is a complex task. The first hurdle is simply identifying which variants contribute to a phenotype. Variants on the same chromosome in close proximity tend to be co-inherited, as they are only segregated in the event of recombination during

crossover. Due to such co-inherited genetic "blocks", termed haplotypes, it must be deciphered whether a variant which is associated with disease actually plays a causative role or is simply co-inherited with an actual causal variant (Sazonovs and Barrett, 2018). Such co-inherited single nucleotide variants are termed marker variants (Gabriel et al., 2002). Although these variants cannot be used to understand the molecular mechanisms underlying a disease phenotype, they can play a useful diagnostic and/or prognostic role. Notably, haplotypes are population specific and have been catalogued by the HAPMAP project (Gabriel et al., 2002; Goldstein and Cavalleri, 2005).

Two major approaches exist to identifying variants with disease associations; both with different strengths and weaknesses. These are linkage or cosegregation studies (Ott et al., 2015), and association studies. Historically, the first genotype-disease associations were discovered through linkage studies (Bodmer and Bonilla, 2008), with causal links established between human leukocyte antigen (HLA) alleles and a number of diseases, including Hodgkin's disease, ankylosing spondylitis, and hemochromatosis (Bodmer, 1973; Feder et al., 1996). Here, variants which cosegregate with a disease phenotype are identified as likely causal variants. Since the first discoveries, this method has been able to identify a number of variant-disease associations, including several disease-associated variants in the titin gene (Seidman and Seidman, 2011). Its limitations are that for cosegregation to be detected, information on family members with and without the disease must be available. Additionally, this method is most applicable to cases with high disease penetrance, although models have been developed to deal with lower penetrance (Ott et al., 2015). This method cannot be applied to *de novo* variants or somatic variants, as these are not inherited.

Association studies instead compare the incidence of variants in case and control groups, where individuals in both groups are unrelated and cases and controls are matched for confounding factors (i.e. ethnicity) (Lewis and Knight, 2012; Tsao and Florez, 2007). The assumption is made that variants associated with disease will be enriched in the case group over the control group. A number of approaches exist to calculating this enrichment, however, most commonly a Fisher exact test is used, and the results subjected to correction for multiple testing. Moreover, it is generally assumed that the impact of variants is additive, and thus each variant can be treated as an independent variable. Using this approach, provided the cohort of cases and controls is large enough, it is possible to detect variants with incomplete penetrance, such as those which may be associated with complex

disease (Lewis and Knight, 2012). However, if causes of a disease are genetically heterogeneous, specific disease-associated variants will not be particularly enriched in the disease cohort, and are thus likely to go undetected.

Both of the above approaches can be combined with a hypothesis-driven or a hypothesis-free approach (Donaldson et al., 2016). In hypothesis-driven approaches, prior information about a disease is used to prioritise particular "candidate" genes for investigation (Patnala et al., 2013). This information is generally retrieved from the literature and/or relevant databases. For example, genes which are associated with a pathway known to be associated with a disease may be prioritised (Patnala et al., 2013). This approach allows for greater statistical power, however has the disadvantage that any completely novel associations cannot be detected. Hypothesis-free approaches, on the other hand, are unbiased but may lack statistical power in detecting disease associations. These approaches generally involve either genome- or exome- wide investigation (Kitsios and Zintzaras, 2009). In contrast, hypothesis-driven approaches are more often restricted to the targeted set of "candidate" genes, thereby reducing costs. However, these may also involve genome-wide investigation, in which prior knowledge is used to weight results (Bakir-Gungor et al., 2014).

Genome-wide association studies and linkage studies have, to date, used primarily microarray chips to identify variants (Sazonovs and Barrett, 2018). These have the advantage of comparatively low cost, and, due to linkage disequilibrium, 2 to 4 times (Donaldson et al., 2016) more single nucleotide polymorphism (SNP) genotypes can be assigned than directly detected, using imputation software (Sazonovs and Barrett, 2018), such as PLINK (Chang et al., 2015; Purcell et al., 2007). The disadvantages here are that any single nucleotide variants (SNVs) which are rare or unique will be absent from these microarrays. With the decreasing cost of NGS technology, many studies are moving towards the use of exome and genome-wide sequencing. Here trade-offs between sequencing depth (and therefore the quality of variant calls), and cost must be considered (Sazonovs and Barrett, 2018).

A number of problem cases still exist. *De novo* variants, defined as variants which are present in a child but not in either parent, are difficult to classify. Such variants have been linked to a number of diseases, including neurodevelopmental disorders. Although these can be detected by sequencing parent-child trios, establishing which *de novo* variants play a causative role in an observed phenotype

can be challenging. This is due to the fact that other individuals with the same condition are unlikely to share the causal variant (Acuna-Hidalgo et al., 2016). A similar problem is observed in the case of somatic cancer variants. As cancers are genetically heterogeneous, although some driver mutations are shared between individuals, a number are thought to be rare or personal and are thus difficult to detect (Hou and Ma, 2014). Additionally, we have already been introduced to the problem case of diseases whose genetic causes are heterogeneous. Furthermore, as stated earlier, it is generally assumed that the impact of variants is additive, however, it is clear that this is an oversimplification. From a biological viewpoint, as gene products, proteins, do not function in isolation but interact with one another as part of a complex system, it seems unlikely that the combined impact of variants will be equal to the sum of its parts. Indeed, a number diseases are known to be digenic (caused by a combination of variants in two genes), such as those catalogued by the DIDA database (Gazzo et al., 2016), and, in Section 1.3.2 we will be introduced to disease-associated titin variants with compound heterozygous inheritance. Moreover, the known phenomenon of compensatory variants provides evidence that variant-phenotype relationships are not always linear (Baresić et al., 2010). The combinatorial impact of variants we discuss here is termed epistasis. This is particularly difficult to detect, as although a number of algorithms have been developed, exhaustive approaches are computationally costly. Moreover, statistical power can be elusive due to the tendency towards large p-values and small numbers of observations (Wei et al., 2014).

These problem cases, with the exception of *de novo* variants and somatic cancer variants, can offer a partial explanation to the phenomenon of "missing heritability", otherwise termed the "dark matter" of genome-wide association (Manolio et al., 2009). As already outlined, this phenomenon occurs when a phenotype is observed to be largely heritable, but only a small portion of this heritability can be attributed to known genetic factors. An often used example is human height; this has a heritability of approximately 80 %, yet despite extensive study, only 5 % of this heritability can currently be explained (Visscher, 2008). At the crux of the matter is understanding how much of this missing heritability can be attributed to common variants with, mainly, small effect sizes (coined the 'common variant common disease' hypothesis) or rare variants with large effect sizes (Gibson, 2012). Indeed, known examples suggest both ends of the spectrum are possible: observed odds ratios for the majority of common variant-disease associations are between 1.1 and 1.4, whereas odds ratios

for rare variant-disease associations are frequently greater than 2, or cannot be calculated due to the absence of the rare variant from the control cohort (Bodmer and Bonilla, 2008). In contrast, a handful of well-known cases exist where common variants have a large impact on phenotype, the classic example being that of sickle cell disease (Luzzatto, 2012). These cases appear to occur only where there is an advantage to being a heterozygote; for example, heterozygosity for the sickle cell allele confers resistance to malaria. It is also likely that any small effect sizes of rare variants go undetected by statistical methods. On a practical note, the detection of rare variants with large effect sizes offers more immediate diagnostic and therapeutic value than detecting common variants with small effect sizes. A variant must have a large effect size for the development of a targeted therapeutic to be commercially viable (Manolio et al., 2009). Likewise, variants with small effect sizes have little diagnostic use, as they can only suggest that individuals are slightly more or less likely to develop a disease (Manolio et al., 2009; Reich and Lander, 2001). In this work, we try to understand better the distinction between the characteristics of common and rare variants, and focus, in particular, on deciphering the impact of rare titin variants. We do this by considering the effect of SAVs on protein structure and function, which we feel to be of fundamental importance in untangling the issues discussed here. It should not be forgotten that non-coding variants can play an important role in disease (Zhang and Lupski, 2015); however, the consideration of these is beyond the scope of this work.

1.2 The impact of SAVs on protein structure and function

The twenty amino acids are the building blocks from which proteins are constructed. This gives rise to 190 possible different single amino acid variants (SAVs), where one amino acid is swapped for another, and a lower number (75) of these result from nsSNVs, as not all amino acids can be transformed into one another by changing a single nucleotide of their codon. In reality, the number of possible impacts is much higher, as protein sequences fold into highly complex tertiary structures. Therefore a multitude of factors must be considered in order to understand the effect of SAVs on protein structure and function.

1.2.1 The localisation of SAVs in 3D space

From studies which map disease-associated variants to 3D protein structures, it has become widely established that disease-associated SAVs are enriched in both protein cores and protein interaction interfaces, but depleted on protein surfaces and inter-domain disordered regions (David et al., 2012; de Beer et al., 2013; Engin et al., 2016; Gao et al., 2015; Gong and Blundell, 2010; Gress et al., 2017; Lu et al., 2015; Petukh et al., 2015). Recent work suggests that the apparent enrichment in interaction interfaces could be due to a bias in studies towards disease-associated proteins and their structures (Gress et al., 2017); we address this issue in Chapter 3. There is also evidence to suggest that disease-associated variants impact on post-translational modifications (PTMs) and functional sites more than neutral variants (de Beer et al., 2013; Gao et al., 2015; Gong and Blundell, 2010; Holehouse and Naegle, 2015). Furthermore, it has been shown that disease-associated variants, although not enriched in DNA-binding proteins, are enriched at DNA-binding interfaces, whereas somatic cancer mutations, although not specifically enriched at DNA-binding interfaces, are enriched in DNA-binding proteins (Gress et al., 2017). Beyond the consideration of localisation to specific regions, it has also been shown that "neutral" population variants are more dispersed throughout 3D protein structures, whereas disease-associated variants tend to form clusters (Sivley et al., 2018). Although recurrent somatic cancer variants generally show patterns similar to population variants, significant clustering of these variants is seen in a subset of proteins (Sivley et al., 2018). This observation has been harnessed by a number of algorithms which make use of clustering on 3D structure to detect cancer driver variants (Porta-Pardo et al., 2017). The spatial dispersion of population variants generally reflects avoidance of the core and important functional sites. Indeed, one example is the depletion of population variants near PTMs, with this depletion being particularly marked for PTMs which cluster in sequence space (Reimand et al., 2015).

1.2.2 The physicochemical impact of SAVs

At the physicochemical level, it has been shown that the distribution and impact of amino acid changes due to disease-associated mutations are distinct from those of neutral mutations. Generally, disease-associated mutations result in more drastic changes in hydrophobicity, charge, size, shape,

and hydrogen bonding networks, whereas neutral mutations more frequently conserve properties of the wild-type (WT) (Alexov and Sternberg, 2013; de Beer et al., 2013; Gong and Blundell, 2010; Petukh et al., 2015). Changes in charge can disrupt salt bridges, and the introduction of charged or hydrophilic residues to the core of a protein can result in structural destabilisation (see Fig. 3.2a). Furthermore, large differences in the size and/or shape of a buried residue will either result in steric clashes, or the formation of undesirable structural voids (Al-Numair and Martin, 2013). More specifically, disease-associated mutations are enriched in changes from the wild-type amino acids cysteine, tryptophan and glycine (Petukh et al., 2015; Vitkup et al., 2003). These play important structural roles, with cysteine able to form disulfide bonds, tryptophan being the largest amino acid, involved in hydrophobic, cation-pi and pi-stacking interactions, and glycine the smallest, most flexible amino acid (Gao et al., 2015; Makwana and Mahalakshmi, 2015). Disease-associated mutant amino acids are most enriched in mutations to proline and cysteine (Gao et al., 2015). Mutations to cysteine may result in the formation of unwanted, disruptive, disulfide bridges, whereas the rigid proline is known as a helix breaker and can also disrupt hydrogen bonding in turns. Interestingly, disease-associated mutations to proline are only enriched in ordered secondary structural elements, and not in disordered coil regions (Gao et al., 2015). Consistent with disease-associated mutations having a greater impact on protein structure and function, are observations that disease-associated variants result in greater changes to the folding and binding free energies of proteins than neutral variants do. These conclusions have been made possible by experimental data available from the ProTherm (Gromiha and Sarai, 2010) and SkemPi (Jankauskaite et al., 2018) databases. However, it must be considered that these data are likely biased towards the study of disease-associated variants and variants of interest to protein engineering.

1.2.3 The impact of SAVs on functional sites

The impact of variants can also be better understood by considering the local context around the affected residue. Mutations at or close to functional sites (e.g. metal-coordinating sites, such as the zinc-coordinating site in p53 that resides near the p53-TP53BP2 interface (Fig. 1.1b)) (Gorina and Pavletich, 1996), or post-translational modification sites can disrupt these (Reimand et al., 2015). Such mutations can be either gain of function, i.e. result in greater enzymatic activity, or loss of

function. Disease mutations which disrupt PTMs can result in the rewiring of signalling networks (Reimand et al., 2015) and can impact on protein-protein interactions (Lu et al., 2016a; Petschnigg et al., 2017).

1.2.4 The impact of SAVs on protein flexibility and disorder

It has been shown that population variants are enriched in disordered inter-domain protein regions, whereas disease-associated variants, in contrast, are more enriched in ordered domain regions (Lu et al., 2015). This is likely due to the fact that disordered regions are subject to less structural and functional constraint. However, this is an oversimplification, as intrinsically disordered regions can play an important biological role; for example in signalling, disorder to order transitions, protein interactions and in the transmission of allosteric signals. Furthermore, such proteins are often tightly regulated, with functions which can be tuned via PTMs (Babu et al., 2011; Dunker et al., 2015; Fong and Panchenko, 2010). Beyond pure intrinsically disordered regions, the degree of regional flexibility is particularly important to a protein's dynamics in the event of conformational transitions. Indeed, there are examples of disease-associated mutations which impact on protein flexibility, thereby altering the equilibrium between different conformers of a protein. A particular case of this is demonstrated by mutations in allosteric proteins (Pandini et al., 2012; Weinkam et al., 2013), such as somatic cancer mutations in phosphofructokinase-1 (Webb et al., 2015), which can impact on the conformational equilibrium between effector bound and unbound forms. Moreover, the impact may not result in a simple re-weighting of conformations observed in the wild-type, but can also result in the exploration of new conformations (Weinkam et al., 2013).

1.2.5 From the atomistic to the macroscopic level

The atomistic impact of variants on protein structures will ultimately exert its effect at the macroscopic system level. A network of protein-protein, protein-nucleotide and protein-ligand interactions make up this system. Variants which localise to the core of a protein may disrupt tertiary structure (Fig. 1.1a), and have the potential to abrogate association with all interaction partners (node removal), whereas mutations which localise to interaction interfaces may disrupt specific interactions (edges in the network, Fig. 1.1c) (Jubb et al., 2017; Laskowski and Thornton, 2008; Yi et al., 2017). Variants

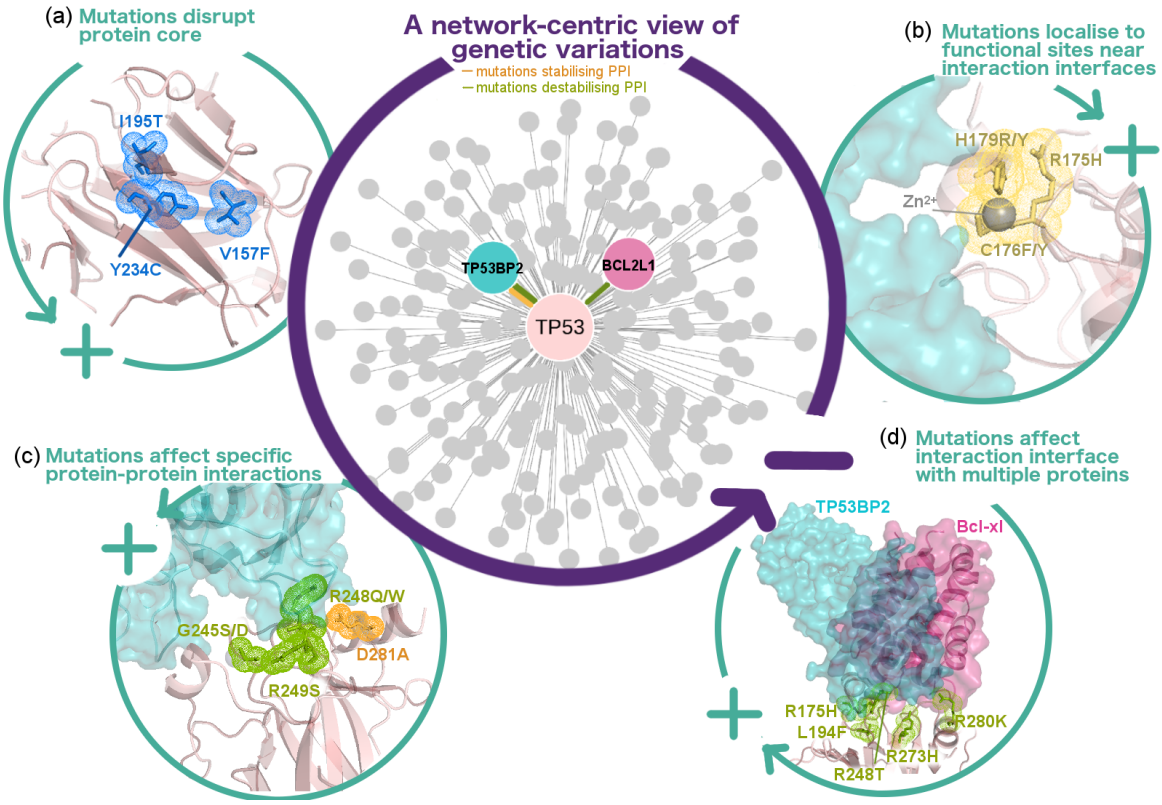

Fig. 1.1 Molecular mechanisms of genetic variants which affect PPIs. Here we take the example of the human p53 protein (encoded by *TP53* gene), first zooming out ("-" sign) to look at its directly interacting proteins (purple circle), then zooming in ("+" sign) to study atomistic details of the mutated residues (green circles). The same colour code is used in both the PPI network and the structures, which show two structural complexes of p53 (in salmon), one with the pro-apoptotic protein p53 binding protein-2 (in cyan, encoded by *TP53BP2*, PDB 1ycs (Gorina and Pavletich, 1996)) and the other with the anti-apoptotic protein Bcl-xl (in violet-red, encoded by *BCL2L1*, PDB 2mej (Follis et al., 2014)). Four categories of mutations were shown here (a-d, see main text), with the wild-type residues highlighted on the structures. All mutations have been observed in cancers, curated from both the COSMIC database (v83, http://cancer.sanger.ac.uk/cosmic) (mutations in $\geq$ 100 cases were considered) and the literature (Cho et al., 1994; Nishi et al., 2013; Tomita et al., 2006). Their effects on PPIs were assessed to be either stabilising (in orange) or destabilising (in green) *in silico* (Nishi et al., 2013) or *in vitro* (Cho et al., 1994; Tomita et al., 2006). Structures were visualized and rendered in PyMol (v1.8.2.1). Network was compiled through STRING-db (v10.5, https://string-db.org) (high-confidence interactions, score $>$ 0.95) and visualized in Cytoscape.

which destabilise a protein may also result in pathogenicity through loss of specific, for example enzymatic, functions, or the accumulation of toxic aggregates. Examples being the cardiovascular disease predisposing T560M mutation, found in human lipoxygenase A, which disrupts a hydrogen bonding network connecting the mutation to the active site (Schurmann et al., 2011), and the Charcot Marie tooth disease associated L16P mutation in peripheral myelin protein 22, which leads to protein misfolding and aggregation (Sakakura et al., 2011). A subtle difference must be noted between those variants which destabilise the folded form of the protein and those which impact on folding kinetics (Zhang et al., 2012). Conversely, it has also been shown that deleterious missense variants can exert a stabilising effect. For example, the H101Q CLIC2 mutation, associated with X-linked intellectual disability with cardiomegaly, stabilises the CLIC2 protein; thereby preventing conformational changes which are necessary for its transition between soluble and membrane forms (Takano et al., 2012; Witham et al., 2011).

Experiments have shown that the majority (approximately two-thirds) of disease-associated variants perturb PPIs but preserve protein stability (Sahni et al., 2015), contingent with them specifically affecting interaction interfaces. Such interface mutations can alter binding affinity, or, more rarely, result in novel interactions (Sahni et al., 2015). The p53 protein, a known cancer driver, offers examples of both variants which disrupt the core of a protein (Fig. 1.1a) and variants which affect specific interfaces, such as the p53-TP53BP2 interaction interface (Fig. 1.1c) (Cho et al., 1994; Gorina and Pavletich, 1996; Nishi et al., 2013; Tomita et al., 2006).

Cancer driver mutations offer an ideal example to illustrate how such types of genetic variants can affect PPIs. Here, differences in the protein regional enrichment in somatic mutations between subsets of cancer driver genes have been observed. Proteins encoded by oncogenes have been found to be enriched in variants which localise to both protein interaction interfaces and functional sites, while those encoded by tumour suppressor genes have been found to be enriched in variants which are found in the core of a protein (Engin et al., 2016; Stehr et al., 2011). Moreover, Engin et al. (2016) found that the interfaces of tumour suppressors annotated as activating interfaces are enriched in destabilising mutations, in comparison to such interfaces of oncogenes, thereby offering mechanistic insight into the differences between the two classes of cancer drivers.

A single protein may have multiple interfaces; additionally, it may use particular interfaces to interact with multiple proteins. p53, for example, employs different interfaces to bind with TP53BP2 and the anti-apoptotic protein Bcl-xl (Fig. 1.1d) (Follis et al., 2014; Gorina and Pavletich, 1996). In a number of pleiotropic proteins (those associated with multiple biological functions) (Chesmore et al., 2016), mutations associated with distinct diseases have been observed to cluster on different interfaces (Mosca et al., 2015; Wang et al., 2012). On the contrary, different variants which localise to the same PPI interface can result in the same/similar phenotypes (Mosca et al., 2015; Wang et al., 2012). Interestingly, disease-associated mutations which abrogate a greater number of interactions have been correlated with earlier disease onset (Sahni et al., 2015). On the other hand, it has been observed that common variation is depleted in multi-binding promiscuous interfaces (Fornili et al., 2013). Other studies into conformational dynamics have shown protein interfaces to be more rigid; furthermore, those interfaces which harbour disease-associated variants were demonstrated to be even less flexible (Butler et al., 2015). Here it must be taken into consideration that protein interactions are heterogeneous in both structural and physicochemical properties, therefore this trend may not hold for all protein interaction interfaces. Indeed it has been shown that, in some cases, flexibility can play an important role; in particular where partner binding is mediated by a continuous epitope, as well as in interactions involving intrinsically disordered proteins, where disorder-to-order transitions may occur (Jubb et al., 2015). In such cases, it has been demonstrated that mutations which decrease the level of disorder in the unbound conformation or change the propensity for secondary structure formation upon binding, can be pathogenic (Yates and Sternberg, 2013).

Protein interaction interface regions can be further divided into rim regions, which are similar in composition to the rest of the surface, and core regions which are more hydrophobic in nature. Interface core regions have been found to be most highly enriched in disease-causing SAVs, whereas neutral SAVs appear to localise preferentially to rim regions. Hotspot residues, which provide the largest energetic contribution towards binding have, unsurprisingly, been found to be enriched in disease-associated SAVs (David and Sternberg, 2015).

Missense variants may also impact on protein-nucleic acid interaction interfaces. Such mutations can lead to altered gene regulation, translation, RNA editing and DNA replication and repair

(Gommans et al., 2009; Pires and Ascher, 2017; Zhang et al., 2012). This has been particularly noted in the case of cancer, for example, certain p53 mutants demonstrate impaired binding to DNA response elements (Muller and Vousden, 2013). Of note, disruptive mutations may occur on either the interface of the protein or the nucleic acid. For example, it has been found that cancer mutations are enriched in DNA-cohesin binding sites (Katainen et al., 2015).

1.2.6 Proteomics and transcriptomics data

Insights into the behaviour of proteins and related transcripts, in their cellular environment, can be obtained through the use of mass spectrometry and RNAseq technologies. Interestingly, correlations between measured protein abundances and mRNA levels are incomplete, for example, a review by Maier et al. (2009) reports Spearman correlations ranging between 0.45 and 0.75. This discrepancy between transcript levels and protein abundance can be attributed to biological processes, such as those associated with translation, mRNA stability, and protein turnover, in addition to technical error and noise (Maier et al., 2009). Despite these incomplete correlations, it has been shown that both proteins which are highly transcribed, and those which are abundant, evolve at a slower rate. Presumably, as any mutations which result in destabilisation or aggregation will impact on a larger quantity of protein in the crowded cellular environment, leading to greater cellular toxicity. Interestingly such correlations, between protein evolutionary rate and abundance, are strongest for proteins expressed by tissues with a high neuron density (Drummond and Wilke, 2008). This can be attributed to the greater potential for misfolded proteins and aggregates to lead to toxicity, due to the long neuronal life-span over which these can build up. It has also been shown that the surfaces of more abundant proteins have a lower "stickiness" than less abundant proteins (Levy et al., 2012). This property has been investigated using the "stickiness" index, which is based on the propensity of an amino acid to occur at protein-protein interaction interfaces. This suggests that there is a greater need for abundant proteins to avoid non-specific interactions. Furthermore, it has been observed that evolutionary mutations which result in large "stickiness" changes are less common on the surfaces of abundant proteins. Observed correlations between stickiness and abundance are higher for the Escherichia coli proteome than the human and yeast proteomes. This is likely

due to the fact that the compartmentalisation of eukaryotic cells results in non-linear relationships between protein concentration and abundance (Levy et al., 2012).

Recent developments in mass spectrometry-based proteomics now enable the measurement of a number of protein properties, beyond abundance, such as thermal stabilities and half-lives (Leuenberger et al., 2017; Mathieson et al., 2018). Data derived by such techniques have shown that proteins with high thermal stability are more abundant. In light of the discussed correlations of abundance with evolutionary rate, this suggests that more abundant proteins have evolved to become more stable, due to the potentially detrimental effects large quantities of aggregated or misfolded proteins would have on the cell. It has also been shown that essential proteins have higher thermal stability than non-essential proteins, and that stable proteins are enriched in functions distinct to those of unstable proteins. Specifically, the Picotti lab has shown that stable proteins are enriched in ribosomal, RNA-binding and protein biosynthesis processes, whereas unstable proteins are enriched for cofactor and DNA-binding proteins (Leuenberger et al., 2017).

We believe that these measured properties of proteins, in particular their abundance and stability, could give insight into the potential impact of localised missense variants. The correlations of evolutionary rate with this data, suggests that this may be the case; however, to the best of our knowledge, the relationship of these features with the enrichment of missense variants has not been investigated. We explore these possibilities in Chapter 3.

1.3 Titin

1.3.1 Titin structure and function

The giant protein titin spans half a cardiac sarcomere from the Z-disk to the M-line, and plays a pivotal role in sarcomeric function and stability. Notably, the I-band region is characterised by its elastic properties (Gigli et al., 2016) and maintains resting tension, whereas the A-band region has been suggested to act as a blueprint for thick filament assembly, although recent research suggests that a better description is that of a wrapper which restricts the length of the thick filament (Kellermayer et al., 2018). Although titin is known to be involved in numerous interactions, including interactions with the thick filament proteins myosin and myosin binding protein C in the

A band region (Chauveau et al., 2014b), only a handful of these have been structurally characterised. These are titin's interactions with telethonin at the beginning of the Z-disk, and both obscurin and obscurin-like protein at the end of the M-line.

Titin is modular in structure, being comprised primarily of the globular immunoglobulin and fibronectin type-III (Fn3) domains. Notably, titin's immunoglobulin domains are classified as belonging to the intermediate-set (I-set), as they have loop regions which are intermediate in length between variable-set (V-set) and constant-set (C1) domains (Kenny et al., 1999). Ig domains are found throughout the length of titin, whereas Fn3 domains localise only to the A-band region, where, along with the Ig domains they form super-repeat patterns (see Fig. 1.2). These comprise of 7 and 11 domains in regions termed the D-zone and C-zone respectively. The only departures from these patterns are at the beginning and end of the A band. Evolutionary analysis has shown that domains at particular positions within the super-repeat are more closely related to one another than to domains at other positions (Kenny et al., 1999). Interestingly the spacing of the D-zone repeats, at a length of 43 nm, matches the periodicity with which myosin binding protein C interacts with titin, supporting the molecular blueprint hypothesis (Kenny et al., 1999). Additionally, a single pseudokinase domain is found at the beginning of the M-band region, which acts as a mechanosensor (Bogomolovas et al., 2014). Both titin Fn3 and Ig domains share a similar beta sandwich structure with a conserved hydrophobic core (Meyer and Wright, 2013). The major difference between these two domain types is in the topology (see Fig. 1.2). It has also been shown that titin Ig domains have higher mechanical stability than titin Fn3 domains, however, even titin Fn3 domains have a higher mechanical stability than those of the extracellular protein tenascin. Furthermore, domains from the A-band region have been shown to have lower mechanical stability than those from the I-band region, throughout which stability increases from the N to C terminal end. The lower stability of A-band domains is perhaps indicative of the fact these are most likely stabilised by their numerous interactions with the thick filament (Rief et al., 1998). Interestingly, when I-band Ig domains are divided into more stable and less stable domains, particular sequence motifs unique to each group have been found (Garcia et al., 2009). The ability of titin's I-band Ig domains to unfold and refold in response to physiological stretch forces contributes to the spring-like properties of this region (Li and Linke, 2017; Meyer and Wright, 2013). Despite their discussed differences, both titin Fn3 and Ig

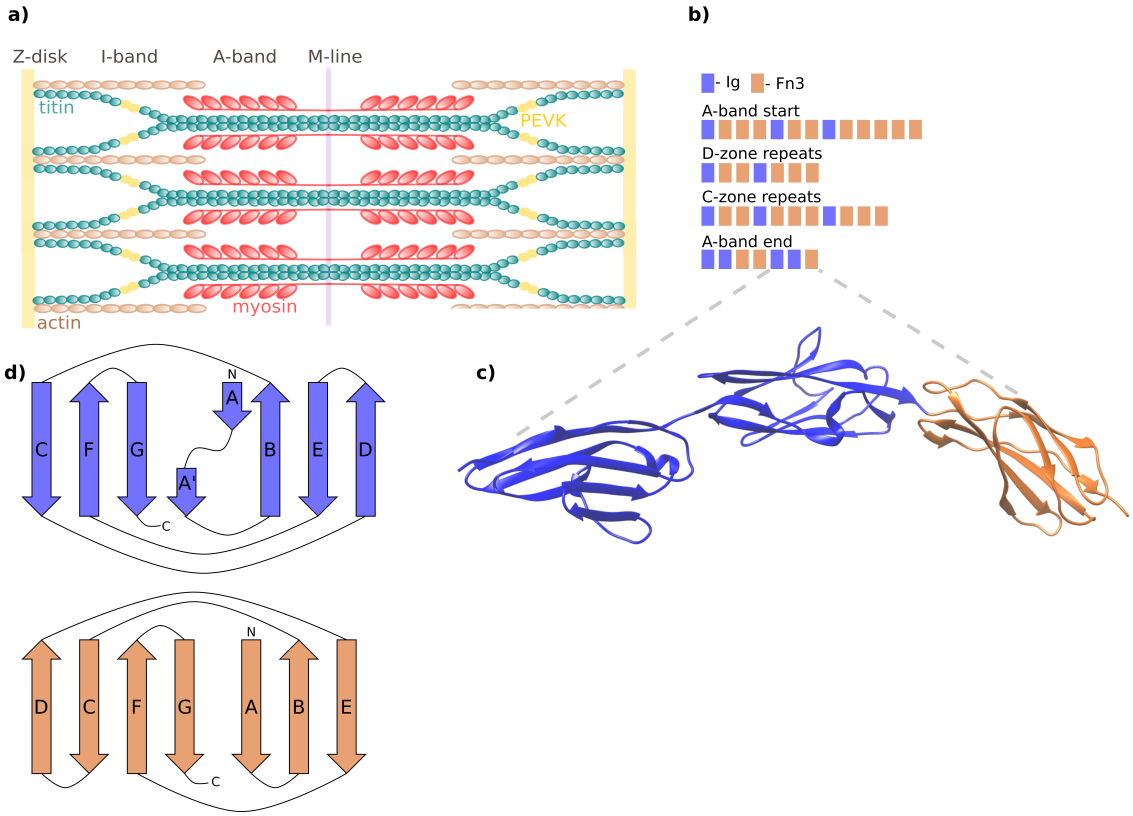

Fig. 1.2 a) Schematic of titin's localisation within the sarcomere; a single titin molecule (green) spans half a sarcomere from the Z-disk to the M-line. The I-band region of titin acts as an elastic spring, and interacts with proteins of the thin filament, whereas the A-band region interacts with the thick filament proteins myosin and myosin binding protein C. b) Titin A-band domain organisation. c) C-terminal domains of the A-band (Ig-158, Ig-159, Fn3-132), PDB structure 2nzi. Structures were visualised using the UCSF Chimera software (Pettersen et al., 2004). d) The topology of titin Ig (blue) and Fn3 (orange) domains.

domains fold independently and are stable at room temperature. Moreover, crystal structures are generally easy to obtain, although it has not yet been possible to obtain a crystal structure for the HMERF hotspot domain Fn3-119 (Meyer and Wright, 2013). In addition to these globular domains, titin contains a flexible PEVK region, so named for its amino acid composition, which is a hotspot for interactors, including tropomyosin and actin (Chauveau et al., 2014b).

A number of different titin isoforms exist, seven of which are documented in the RefSeq database (O'Leary et al., 2016). The longest isoform is the inferred complete (IC) isoform; although this has no known biological relevance, this provides a useful reference point for the positional numbering of other isoforms. The major biologically relevant isoforms are the N2A, N2B, and N2AB variants.

These differ in length due to differential splicing in the I-band region, with the shortest isoform being the cardiac-specific N2B isoform. Specifically, the skeletally expressed N2A isoform contains the N2A exons, the N2B isoform contains the N2B exons and a reduced number of PEVK region exons, and the N2BA isoform contains both the N2A and N2B exons. The Z-disk, A-band, and M-line regions are, in contrast constitutively spliced in. Additionally novex-1 and -2 isoforms are similar to the N2B isoform but contain unique 125 and 192 amino acid stretches respectively. Finally, a unique isoform, called novex-3, contains an alternative terminal out of frame exon, which cannot be mapped to the reference IC isoform (Chauveau et al., 2014b).

1.3.2 Disease-associated titin variants

Due to titin's large size, associations of titin variants with disease phenotypes was, mainly, not feasible until the advent of NGS technology. Since then a number of rare titin variants have been associated with disease (Savarese et al., 2018). Most notably hotspots in the domains Ig-169 and Fn3-119 have emerged for the diseases tibial muscular dystrophy/limb-girdle muscular dystrophy 2J (TMD/LGMD-2J) and hereditary myopathy with early respiratory failure (HMERF) respectively (Chauveau et al., 2014b; Savarese et al., 2016). As well as missense variants a number of truncating variants in titin have been identified, and are now known to be one of the primary causes of dilated cardiomyopathy (DCM) (Schafer et al., 2017). Although truncating variants are also present in approximately 1 % of the general population, it has been found that those variants which lead to the DCM phenotype localise to exons which are constitutively spliced in, as these cannot be "rescued" by differential splicing (Schafer et al., 2017). Moreover, high-resolution cardiac scans have revealed eccentric cardiac remodelling in those nominally "healthy" individuals who possess truncating variants in constitutively spliced in exons (0.5 % of the population), thereby exposing a phenotypic continuum (Schafer et al., 2017). In the last few years, compound heterozygous mutations have been associated with severe paediatric titinopathies, including novel forms of core myopathy with heart disease (Chauveau et al., 2014a). Although only a handful of cases have been characterised, these generally involve the inheritance of a rare titin missense variant in trans with a truncating variant. One proposed hypothesis is that the truncating variant cannot give rise to a functional titin molecule, and thereby unmasks the deleterious nature of the missense variant. A number of variants

have also been identified in the titin kinase (TK) domain, one of which was initially associated with the disease HMERF (Pfeffer et al., 2014), however as this is in linkage with a Fn3-119 hotspot mutation, it is now assumed that this SAV, may at most, be a phenotype modifier. Interestingly, one of the compound heterozygous paediatric titinopathy patients lacks a functional TK domain, due to the possession of a truncating variant, which leads to the loss of almost the entire portion of the C-terminal domain of titin, and a missense W34072R (W260R) TK variant which leads to loss of function (Chauveau et al., 2014a). Please note that here titin variants are notated according to their position in the full length IC isoform with their domain position given in brackets. Including the mutations discussed here, a total of 42 disease-associated missense variants exist with evidence in the literature. These are associated with a variety of myopathies and cardiomyopathies, including hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM) and arrhythmogenic right ventricular cardiomyopathy (ARVC) (Laddach et al., 2017). Additionally, 3 variants, with in-house evidence providing disease associations, can be added to this total. An in-depth discussion of titin-associated pathologies is beyond the scope of this work, for a detailed review of the subject refer to Chauveau et al. (2014b). Suffice to suggest that the impact of titin variants is complex, due to the variety of phenotypic manifestations, and is likely influenced by both the use of different titin isoforms, and distinct cellular environments found in different muscle types (i.e. skeletal and cardiac muscles).

A number of published studies have attempted to elucidate the molecular mechanisms underlying mutations of the Fn3-119 and Ig-169 hotspots. Work by Hedberg et al. (2014) established that the P31709R (P2R), C31712R (C5R), W31729 (W22C), and P25L (P31732L) mutations lead to impaired solubility of the Fn3-119 domain, whereas common population variants in this domain had no such impact. This parallels in-house data which suggests disease-associated Fn3-119 variants can have a spectrum of impacts on the stability of the domain. Similarly, the biophysical impact of the TMD associated missense mutants, 35946P (H50P), L35956P (L60P), and I35947N (I51N) has been studied by Rudloff et al. (2015). In addition, they investigated the impact of an indel, referred to as the FINmaj variant, which results in the sequence change 35927EVTW→VKEK (31EVTW→VKEK). Again their results showed a range of impacts, with the FINmaj and L35956P (L60P) mutants remaining unfolded at room temperature, the 35946P (H50P) mutant remaining folded at room

temperature but demonstrating reduced stability, and the I35947N (I51N) mutant showing only minimal destabilisation. Of the two unfolded mutants, only the FINmaj mutant had a tendency to form aggregates. Similarly, all variants, with the exception of the I35947N (I51N) mutation, had a negative impact on binding to the domain's interaction partner, obscurin. Given these findings, the authors suggest that the I35947N (I51N) mutant may not be causative of TMD, but co-inherited with an actual causal allele. However, to date, no other explanatory allele has been proposed, and it seems unlikely for a non-causal allele, identified by co-segregation, to localise to the structural TMD hotspot.

In addition to mutational hotspots, several isolated mutations have been investigated in the globular Ig and Fn3 domains. The V49M (V54M) Ig-1 mutation, near the beginning of the Z-disk, provides one such example. Here a computational approach was taken and a variety of techniques explored. These included the use of variant impact predictors, protein-protein docking to explore the Ig1-telethonin interface, and molecular dynamics simulations (50 ns). Results from this study suggest that the mutation may result in decreased affinity to telethonin, through destabilisation of the domain and alterations to its secondary structure composition (Kumar et al., 2017).

The majority of variants discussed so far appear to have a destabilising impact, which in some cases has the additional effect of decreasing or abrogating binding-partner affinity. In contrast, the four variants known to be associated with HCM exhibit gain of function effects, resulting in the upregulation of interactions. Specifically, the R740L mutation in the Z-disk increases alpha-actinin binding affinity (Satoh et al., 1999), the S4116Y mutation increases binding affinity to DRAL/FHL2 (Matsumoto et al., 2005) and both the R9744H and R8500H mutations increase the titin/T-cap interaction affinity (Arimura et al., 2009). These mutations are only notated according to their position in the IC isoform as they do not occur within globular domains.

Importantly, hundreds of rare titin missense variants of unknown significance have been identified in disease cohorts. These include patients with HCM (Lopes et al., 2013), peripartum cardiomyopathy (van Spaendonck-Zwarts et al., 2014), DCM (van Spaendonck-Zwarts et al., 2014), and other diseases associated with sudden cardiac death (Campuzano et al., 2015). This highlights the necessity for the development of robust methods to classify rare titin variants.

1.4 Predicting the impact of missense variants

After the analysis of genomic data, typically 100 to 1000 variants (Li et al., 2013) may emerge as suspects for playing a role in a disease phenotype. These may have been identified through enrichment, co-segregation or rare variant/candidate gene analyses, as outlined in Section 1.1. At this stage, it becomes necessary to understand the likely impact of these variants. Here the initial use of computational predictors is desirable, as results from these can be used to further narrow down candidate variants for wet lab analysis. Moreover, computational results can be used to design more targeted wet lab experiments. For example, if a variant is predicted to impact on protein stability, assays for probing this in vitro can be designed. Conversely, if a variant is predicted to impact on protein-protein interaction affinity, wet-lab techniques which probe this will be more appropriate. A number of computational approaches to predicting the impact of SAVs exist, which we will discuss here.

Three main ingredients are required for the creation of variant impact predictors. These are 1) benchmark data which comprises of variants with known or 'labelled' impact; 2) input features; 3) a method or algorithm which can transform these features into a prediction. We will discuss each of these ingredients in turn. Finally, we will draw our attention to recent experimental validations of variant impact predictors, which suggest that these computational methods require further improvement.

1.4.1 Benchmark datasets

The majority of benchmark datasets label variants with known disease associations as deleterious, and "common" population variants as neutral. A summary of the major benchmark datasets is given in Table 1.1. Deleterious variants are generally obtained from the UniProt (The UniProt Consortium, 2018) and ClinVar (Landrum et al., 2016) databases. Neutral population variants are obtained from either UniProt (The UniProt Consortium, 2018) or dbSNP (Sherry et al., 2001), where neutral variants have generally been defined as those with a minor allele frequency (MAF) > 0.01 or 0.05 (as data for more individuals have become available the commonly used thresholds have decreased) (Grimm et al., 2015; Ponzoni and Bahar, 2018). Such neutral variants in dbSNP, at the time

these data sets were constructed, originated from the 1000 genomes project. This contains data for 2502 nominally healthy individuals. As these datasets have been constructed at different time points and contain only partially overlapping data, one strategy has been to amalgamate unique variants to create larger benchmark datasets (Ponzoni and Bahar, 2018). One issue here is that some variants which were previously believed to play a causal role in disease phenotypes are now believed to be phenotypically neutral. This initial misclassification can be attributed to coinheritance (linkage) with a variant which actually plays a causal role. An example of this phenomenon is the titin kinase variant discussed in Section 1.3.2, initially believed to play a causal role in HMERF. Now it is known that this variant is simply coinherited with the causal variant, which localises to the HMERF hotspot domain Fn3-119. Another issue is that a number of variants were previously believed to be common, due to the undersampling of particular populations. Now that genetic data exists for a greater number of individuals, this no longer proves to be the case. Finally, a large assumption the vast majority of these datasets make is that common variants are non-pathogenic and have similar properties to rare non-pathogenic variants. As discussed in Section 1.1, the distinction between rare and common variants is not fully understood. Indeed, it has been highlighted by Li et al. (2013) that rare neutral variants are more difficult to separate from disease-associated variants than common neutral variants are. Moreover, distinguishing disease-associated rare variants from rare neutral variants is closer to the real task at hand, as most identified candidate variants are rare.

An alternative approach is to predict whether a variant will impact on protein molecular function, rather than establishing whether it is deleterious or not. This approach was taken in the construction of the benchmark dataset for the SNAP2 predictor (Hecht et al., 2015). A number of sources were used in the creation of this dataset, including the Protein Mutant Database (Kawabata et al., 1999). This database extracts data from the literature on mutant proteins and, amongst other annotations, records whether a change in activity or stability is reported. The SNAP2 dataset also takes information from enzymes which have the same EC classification and > 40 % sequence identity. Here it is assumed that any amino acid differences between pairwise alignments of such enzymes constitute functionally neutral mutations (Hecht et al., 2015).

In a similar vein, a number of predictors aim to decipher whether variants are destabilising. Here benchmark datasets are extracted from the ProTherm (Gromiha and Sarai, 2010) and SkemPi

Name	Deleterious subset	Neutral subset
HumVar (Adzhubei et al., 2010)	22,196 disease-associated mutations from UniProtKB.	21,119 common nsSNPs (MAF $>$ 0.01) from UniProtKB annotated as non-damaging.
ExoVar (Li et al., 2013)	5,340 disease-associated alleles with known impact on molecular function from UniProtKB	4,752 rare alleles (MAF $<$ 0.01) from dbSNP build 131, with no known disease association and at least 1 homozygous genotype in the 1000 genomes project.
VariBench (Nair and Vihinen, 2013)	19,335 missense variants from the PhenCode database	21,170 nsSNPs with allele frequency $<$ 0.01 and chromosome sample count $>$ 49 from dbSNP build 131, with no known disease-associations.
predictSNP (Bendl et al., 2014)	19,800 disease-associated variants from a large number of databases (SwissProt, HGMD, HumVar, Humsavar, dbSNP, PhenCode, IDbases, 16 locus-specific databases).	24,082 neutral variants from a large number of databases (SwissProt, HGMD, HumVar, Humsavar, dbSNP, PhenCode, IDbases, 16 locus-specific databases).
SwissVar (Mottaz et al., 2010)	36,440 disease-associated variants from the UniProtKB/Swiss-prot database.	40,193 variants with no known disease-associations from the UniProtKB/Swiss-prot database.
Rapsody dataset (Ponzoni and Bahar, 2018)	Non-redundant union of deleterious SAVs which can be mapped to a PDB structure from the HumVar, ExoVar, VariBenchSelected, predictSNPSelected and SwissVarSelected datasets.	Non-redundant union of neutral SAVs which can be mapped to a PDB structure from the HumVar, ExoVar, VariBenchSelected, predictSNPSelected and SwissVarSelected datasets.

Table 1.1 Benchmark datasets used in variant impact prediction. Adapted from (Grimm et al., 2015).

(Jankauskaite et al., 2018) databases. The ProTherm database contains data which describes the impact of missense variants on the stability of monomeric proteins, whereas the SkemPi database has analogous information for protein complexes. It must be considered that all such datasets, which amalgamate experimental data from different sources, are likely biased in coverage towards commonly studied proteins and disease-associated variants.

The recent increase in experimental saturation mutagenesis datasets has enabled the use of these as benchmark datasets. These have the advantage that the impacts of each mutant are experimentally validated and associated with a magnitude. Furthermore, the impact of a range of different mutations at a particular protein position can be explored. Here the proviso is that mutations may impact on different functions from those which are measured (Gray et al., 2018).

In selecting benchmark datasets it is important that two types of circularity are avoided, as these can falsely inflate the performance of a predictor (Grimm et al., 2015). Type one circularity arises when variants in the training set and testing set overlap. This is most frequently a problem in the creation of meta-predictors; those predictors which rely on the combined outcome of a number of other predictors to reach their final prediction. Here it is essential that data to train the individual

predictors is not used to assess the meta-predictor. The SwissVarSelected, VariBenchSelected and predictSNPSelected datasets have been particularly designed to overcome this problem, as they consist of variants which are not part of the major previously created benchmark datasets (Grimm et al., 2015).

Type two circularity arises when all benchmark variants which localise to particular protein belong to the same class (i.e. all variants are either classed as neutral or deleterious) (Grimm et al., 2015). This is problematic as a predictor can achieve good performance simply by classifying all benchmark variants which localise to a particular protein as deleterious or neutral. However, in reality, it is likely that variants will fall into both classes. One solution has been to create benchmark datasets only using proteins to which both classes of variants localise (Ponzoni and Bahar, 2018). Unfortunately, this does little to ameliorate the problem if the ratio of variants from each class, which are found in a particular protein, is still severely biased. This problem is further exacerbated by properties intrinsic to the datasets. Once mapped to available structures, almost all of these contain a larger proportion of disease-associated than neutral variants (Ponzoni and Bahar, 2018); this is unlikely to reflect the true distributions of deleterious and neutral SAVs.

1.4.2 Features and methods

Prediction of the impact of missense variants rests on the premise that deleterious variants exhibit properties which are distinct from those of neutral variants. Therefore, a step preliminary to prediction is to uncover such segregating properties, which are referred to as features. We have seen that disease-associated variants differ in their impact on protein structure in a number of ways. However, predicting variant impact from structure alone poses several problems. Firstly, the structural coverage of the proteome is incomplete (see Fig. 1.3), therefore limiting the applicability of these methods. Secondly, due to the diverse ways in which variants can impact on protein structure, a "one size fits all" approach cannot be used (Tang and Thomas, 2016). These problems have led to the more extensive use and development of sequence-based predictors. Such predictors rely primarily on evolutionary information (see Fig. 1.3), based on the notion that mutations at structurally and functionally important sites will be subject to negative selection. Although these have had much success, several issues have been encountered. One problem is that functionally important sites are

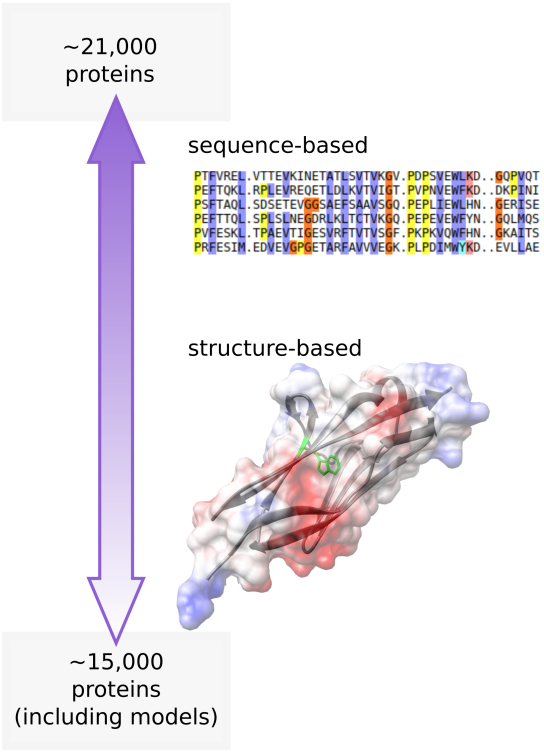

Fig. 1.3 Sequence-based variant impact predictors achieve high coverage of the human proteome (~21,000 proteins) and are most commonly based on evolutionary methods. Structure-based predictors explicitly take the physicochemical environment of a mutation into account; the coverage of such predictors (~15,000 proteins) has been greatly expanded by homology modelling (statistics taken from Bienert et al. (2017) for the human proteome). PDB structure 2y9r is depicted, as visualised by the UCSF Chimera software (Pettersen et al., 2004).

not always conserved between homologs, as paralogs can specifically evolve to perform different functions. This is particularly the case for certain families of enzymes, where homologs catalyse the same class of reaction but the exact substrates differ (Das et al., 2015; Furnham et al., 2016; Stone and Sidow, 2005). Therefore substrate specificity determining residues are not conserved throughout the class. Another problem encountered is that of compensated pathogenic deviations. Here, a variant which is pathogenic in one species is present as the wild-type residue in another species, where its effects are ameliorated by a compensatory mutation. As the mutated residue type is present at the specific position in the alignment, its pathogenic impact is unlikely to be predicted. This is of considerable concern as at least 4 % of human Mendelian disease-associated variants are known to be compensated pathogenic deviations (Azevedo et al., 2016). Moreover, even though the best performing sequence- and structure-based methods have comparable success, sequence-based methods offer little insight into potential mechanisms underlying variant pathogenicity. In the following sections, we will summarise key aspects of both sequence and structure-based approaches. Representative predictors which fall into both classes are summarised in Table 1.2.

Name	Key features	Predictive output	Model type
SIFT (Ng and Henikoff, 2001)	sequence-based predictor (evolutionary information)	deleteriousness	probabilistic
MAPP (Stone and Sidow, 2005)	sequence-based predictor (physicochemical and evolutionary information)	deleteriousness	probabilistic
PolyPhen-2 (Adzhubei et al., 2010)	sequence- and structure-based predictor	deleteriousness	probabilistic
SNAP2 (Hecht et al., 2015)	sequence- and structure-based	functional impact	neural-network
mCSM (Pires et al., 2014b)	structure-based predictor	$\Delta\Delta G$	SVM
HoTMuSiC (Pucci et al., 2016)	structure-based predictor	ΔTm	statistics-based potentials
RAPSODY (Ponzoni and Bahar, 2018)	sequence-, structure- and dynamics-based predictor	deleteriousness	random forests
REVEL (Ioannidis et al., 2016)	meta-predictor	deleteriousness	random forests

Table 1.2 Representative methods for variant impact prediction.

Sequence-based approaches

Sequence-based approaches initially used substitution matrices derived from homologous proteins, such as the well known BLOSUM62 matrix (Henikoff and Henikoff, 1992). Although these give a rough idea of the likely impact of an amino acid substitution, it soon became clear that a residue's

propensity for deleterious mutations is also highly position specific. This led to the development of predictors such as PolyPhen (Sunyaev et al., 2001) and SIFT (Ng and Henikoff, 2001), which derive evolutionary features from multiple sequence alignments of the query protein and homologs. These are the foundations on which the majority of sequence-based predictors are built, with differences lying in the construction of the alignment (generally BLAST or HMM-based), the weighting of each position in the alignment and the calculation of a score which reflects the likelihood of a specific mutation (Tang and Thomas, 2016). More recent developments attempt to tackle the issue of paralogs obscuring conservation, by either filtering sequence alignments to retain only orthologs (Thomas and Kejariwal, 2004), or by using algorithms to detect whether subfamilies diverge (Reva et al., 2011). Of particular interest is the recently developed predictor EVmutation, which creates a global probabilistic model which takes epistatic effects into account, and is able to predict the impact of combinations of mutations (Hopf et al., 2017).

In addition to evolutionary-based features, physicochemical sequence-based properties, such as amino acid hydropathy, polarity, charge and side-chain volume, can be calculated; with an early example being the use of the Grantham scale (Grantham, 1974). This physicochemical approach is exemplified by the MAPP algorithm (Stone and Sidow, 2005) which elegantly combines evolutionary and physicochemical information, with the key step being the calculation of a weighted matrix of physicochemical properties from a multiple sequence alignment, followed by the comparison of this matrix to those properties calculated for the mutant.

Structure-based approaches

Structure-based approaches fall into two major classes; those which use potential energy functions to predict the impact of mutations and those which combine features calculated from 3D protein structure using a machine learning-based approach.

Potential energy functions used by the first class of predictors can be either statistics-based, physics-based or a combination of the two. The Fold-X software (Schymkowitz et al., 2005), for example, uses a physics-based potential, PopMusic (Dehouck et al., 2009) a statistics-based potential and Rosetta (Kellogg et al., 2011) a combined physics and statistics-based potential. An early approach from the Blundell lab, SDM, combines evolutionary and structural information, in its potential energy

functions, by using structural environment-specific substitution tables (Pandurangan et al., 2017). In comparison, physics-based potentials include terms to account for properties such as van der Waals forces, electrostatics, solvation energy, hydrogen bonds, and steric clashes. The weights of such functions may be knowledge-based or learnt from experimental data, for example, FOLDEF (a method based on the FOLD-X energy function) weights energy terms using empirical data from protein engineering experiments (Guerois et al., 2002). Most approaches, which explicitly model a mutation, assume a fixed backbone, although some, such as Rosetta allow for the sampling of backbone conformations (Kellogg et al., 2011). The vast majority of these methods result in a prediction of the change in Gibbs free energy upon mutation, however, the predictor HoTMuSiC uses statistics-based potentials to predict changes in melting temperature (Pucci et al., 2016). This is argued to be a more difficult task than predicting the impact of a mutation on the Gibbs free energy, as it requires an understanding of how this quantity ($\Delta\Delta G$) varies with temperature.

Those algorithms which instead include structural features can either result in a prediction of deleteriousness or impact on stability. PolyPhen-2, for example, incorporates the structure-based features B factor and solvent accessibility, in its prediction of variant deleteriousness (Adzhubei et al., 2013, 2010). The SAAPdb pipeline uses a more extensive range of structural features, including the analysis of structural clashes and the formation of structural voids (where a larger amino acid is replaced by a smaller amino acid) (Hurst et al., 2009). Other structural information used by a number of predictors, includes annotations describing the 3D location of features such as structural domains, binding sites, post-translational modifications and disulphide bridges (Venselaar et al., 2010). These are generally retrieved from databases such as UniProt. When protein structure is available, the predictor I-mutant uses the residue composition within a 9 Å sphere of the mutated residue (Capriotti et al., 2005). The predictor mCSM, in comparison, uses a novel graph-based method to extract atomistic distance patterns from the WT structure, upon which its prediction is based (Pires et al., 2014b). This approach has now been extended to assessing the impact of variants on protein-protein, protein-ligand, and protein-nucleotide interactions. Earlier graph-based approaches include the predictor BONGO (Cheng et al., 2008), which uses a weighted residue-residue interaction network to infer the essentiality of residues and predict the impact of mutations.

Importantly the applicability of these algorithms can be increased through the use of homology modelling. Indeed a handful of predictors, such as HOPE (Venselaar et al., 2010) and VIPUR (Baugh et al., 2016) incorporate this process in their pipeline.

Protein dynamics and flexibility

As discussed in Section 1.2.4, variants can impact on protein flexibility and dynamics. Beyond the incorporation of crystallographic B-factors (Adzhubei et al., 2013) and sequence-based metrics of flexibility, the use of such features in the prediction of the impact of missense variants has been underexplored. One of the few methods to model protein flexibility is Vipur (Baugh et al., 2016), which uses the Rosetta software to model backbone rearrangements. Recently, work from the Bahar lab has shown that incorporating structural dynamics information, derived from elastic network models, can improve the accuracy of prediction (Ponzoni and Bahar, 2018). However, as the dynamic features are based solely on Cα network models, the chemical nature of the change is not modelled. Thus this dynamic information only allows discrimination between positions which are more or less likely to harbour deleterious mutations.

Atomistic molecular dynamics has been used to assess small numbers of "case study" variants. This approach has been applied, in particular, to cases where variants are proposed to impact on allosteric communication (Carluccio et al., 2013). Generally, less computationally expensive predictive methods are employed to prioritise several variants for further study using molecular dynamics techniques. Such molecular dynamics-based investigations have found that potentially deleterious variants generally result in increased flexibility, radius of gyration and solvent accessible surface area (Doss and Nagasundaram, 2012; Kumar and Purohit, 2014; Pires et al., 2017); although particular mutations, such as the V617F JAK2 SAV, have been shown to confer a rigidifying effect. Generally, changes in hydrogen bonding networks have also been observed (Doss and Nagasundaram, 2012; Kumar and Purohit, 2014; Pires et al., 2017). One study performed atomistic molecular dynamics to assess the impact of 57 mutations on the TGFBR2 kinase domain, however, here simulations were limited to two nanoseconds, indicating that structural refinement rather than an exploration of dynamics was achieved (Zimmermann et al., 2017). A few larger scale studies exist which have used alchemical free energy simulations (Seeliger and de Groot, 2010) and free energy perturbation

methods (Steinbrecher et al., 2017) to predict the impact of mutations on the free energy of protein folding, however, differences in protein conformational equilibrium and flexibility have not been explored using MD simulations on this scale. Despite the current limited use of molecular dynamics simulations in the prediction of the impact of missense variants, it has been proposed that these hold promise for assessing variants of unknown impact (Oliver et al., 2016).

Functional and network features

In addition to sequence and structure-based approaches, some success has resulted from using functional and protein-protein interaction network information as predictive features. SuSPect, for example, uses the protein-protein interaction network degree centrality of the protein a variant localises to, to predict the variant's impact (Yates et al., 2014). This is based on the fact that disease-associated variants are believed to localise to proteins with more interaction partners - those proteins which are hubs in the network. Caution must be taken, as the high centrality of disease proteins in protein-protein interaction networks may result from bias due to the fact that they are more frequently studied (Vidal et al., 2011). Other approaches use functional annotations, such as Gene Ontology terms (Capriotti et al., 2013). It is important to note that, although these functional and network features provide additional information in the prediction of variant impact, such protein-level information may lead to an increased probability of type 2 circularity (as discussed in Section 1.4.1).

1.4.3 Combining features and methods

Unless features themselves constitute a predictive value, as is the case for a number of the early evolutionary methods (i.e. the PSIC score of PolyPhen (Sunyaev et al., 2001)), these must be transformed into a prediction of impact. This can be done using either empirical scoring functions or machine learning-based methods. Empirical scoring functions rely on knowledge of the system and can be less reliable if this is incomplete. Machine learning-based methods, in contrast, can infer their functional form from the data and are able to predict non-linear relationships (Wójcikowski et al., 2017). It has been argued that machine learning techniques can be "black boxes" which do not allow insight into how predictions are arrived at, however, the majority of classical machine learning

algorithms allow information to be extracted about the relative importance of different features for making predictions. This may allow for insights which would not otherwise be accessible. A number of machine learning algorithms have been applied to SAV impact prediction, including logistic regression (Baugh et al., 2016), support vector machines (Pires et al., 2014b), random forests (Ponzoni and Bahar, 2018) and neural networks (Ancien et al., 2018).

With the large number of predictors now available, predictors can be combined to form meta-predictors. Again these may be combined empirically, or the results from these predictors themselves input into a machine learning algorithm. Examples of such predictors include Condel (González-Pérez and López-Bigas, 2011), CADD (Kircher et al., 2014) and REVEL (Ioannidis et al., 2016). Although such meta-predictors have been shown to generally out-perform individual predictors, this performance assessment can be inflated by type 1 circularity, as discussed in Section 1.4.1.

1.4.4 Success of predictors

Although the majority of predictors have been shown to perform well on benchmark datasets, a number of recent studies have suggested that predictions do not map well to the actual consequences of missense variants. A seminal study by Miosge et al. (2015) bred mice to homozygosity for 33 potentially disruptive *de novo* single nucleotide variants (30 missense and 3 truncating) in 23 essential immune system genes. Knockout mice for these 23 genes display clear phenotypic readouts. Although only 4 of 30 missense variants resulted in an altered phenotype, 20 of the variants were predicted to be deleterious by the majority of computational predictors used, which included PolyPhen-2, Sift, and CADD. As the authors considered the possibility of phenotypic masking and/or compensation they went on to investigate the in-vitro impact of all possible p53 variants on its transactivation activity. Again, a large proportion of variants predicted to be deleterious (42 % PolyPhen, 45 % CADD) had little impact on p53-promoted transcription. Similarly, Andersen et al. (2017) identified 11 rare variants in 6 genes associated with the interferon induction pathway in patients with herpes simplex encephalitis. Using a HEK293T cell-based system deficient in the protein of interest, they reconstituted with the wild-type or mutant protein by transient transfection, and induction of beta-interferon was assessed using a co-transfected reporter plasmid. An analogous assay was used for variants in the IFNLR1 gene; here naturally occurring variants from the 1000

genomes project were investigated. The results showed that only a single variant from the herpes simplex cohort reduced interferon induction, similarly only one of the 1000 genomes project variants lowered activity of IFNLR1, which still remained at $\sim$ 50 %. However, the majority of computational predictions gave overestimates of the number of deleterious variants (Gray et al., 2018). These results, taken together, indicate that a large portion of computational predictors may lack specificity. This lack of specificity could arise from the fact that predictors are generally trained to separate common from disease-causing variants, and therefore struggle to separate rare deleterious variants from rare neutral variants.

Furthermore, the fact that computational approaches struggle to predict the impact of particular mutations is highlighted by experimental saturation mutagenesis studies (Baugh et al., 2016). In one such study positions in the LacI protein were classified as toggle or rheostat positions (Miller et al., 2017). Here toggle positions are defined as those at which the impact of a mutation is binary, whereas at rheostat positions a gradient in variant impact exists. It was found that, whilst predictors performed well at toggle positions, they were poorly able to distinguish between rheostat positions. Another recent study attempted to use saturation mutagenesis data to train their predictor (Baugh et al., 2016). Although they achieved better performance than other methods on such data, they noted poor performance for three of the twelve studied datasets. No rationale underlying this poor performance could be found.

These facts suggest that, despite the plethora of predictive methods which exist, it is imperative that these methods are improved, in order to better reflect experimental data. We believe this improvement can be brought about by the development and use of better benchmark datasets, and a more comprehensive understanding of which features segregate deleterious from neutral variants, as opposed to deleterious from common variants.

1.5 Contributions of this thesis

In light of the information reviewed here, we sought to contribute towards improving the impact prediction of missense variants. We approached this problem by trying to reach a better understanding of which features can be used to distinguish between neutral and deleterious variants on three

levels, as illustrated in Fig. 1.4. Specifically, we explore protein-level properties, including function, expression, stability and abundance, in Chapter 3, the structural localisation of variants in Chapters 2-4, and the physicochemical impact of variants, including their effect on protein dynamics, in Chapter 4. We focus, in particular, on the titin protein, as variants which localise to this protein can be especially difficult to classify. Additionally, recent NGS studies have uncovered a large number of titin variants of unknown significance (see Section 1.3.2). It must be noted that all work presented in this thesis is associated with human genes/proteins.

The first step towards a better understanding of the impact of variants is their correct and comprehensive annotation; an instruction manual must include a catalogue of its parts. To this end, we developed a web application TITINdb, which we present in Chapter 2. This contains comprehensive annotation of titin variants from the literature, the 1000 genomes project (Auton et al., 2015) and the gnomAD database (Lek et al., 2016). The creation of this resource involved the large-scale modelling of titin domains. This greatly increased the structural coverage of titin and enabled the application of both sequence and structure-based methods to the impact prediction of titin missense variants. All these results are available through the TITINdb interface (http://fraternalilab.kcl.ac.uk/TITINdb/), along with the results of *in silico* saturation mutagenesis.

From our application of currently available impact predictors to titin variants, in Chapter 2, it becomes clear the results do not agree with one another. Moreover, the large number of rare variants which are predicted to be deleterious suggests that these predictors may lack specificity; in agreement with the literature discussed in Section 1.4.4. This motivated us to reach an improved understanding of the characteristics of missense variants in health and disease, and led us to perform a multidimensional analysis of the properties of population variants from the gnomAD database (Lek et al., 2016), somatic cancer variants from the COSMIC database (Forbes et al., 2015) and germline disease-associated variants from the ClinVar database (Landrum et al., 2016). Throughout this analysis, we further segregated population variants into common and rare subsets, in order to better understand the distinction between these. Uniquely, we integrated our analysis with available transcriptomic and proteomics data. Clear differences between the datasets emerge; these include observations which build upon those already reported in the literature but derived using more extensive data, in addition to novel observations. The results of this work are presented in

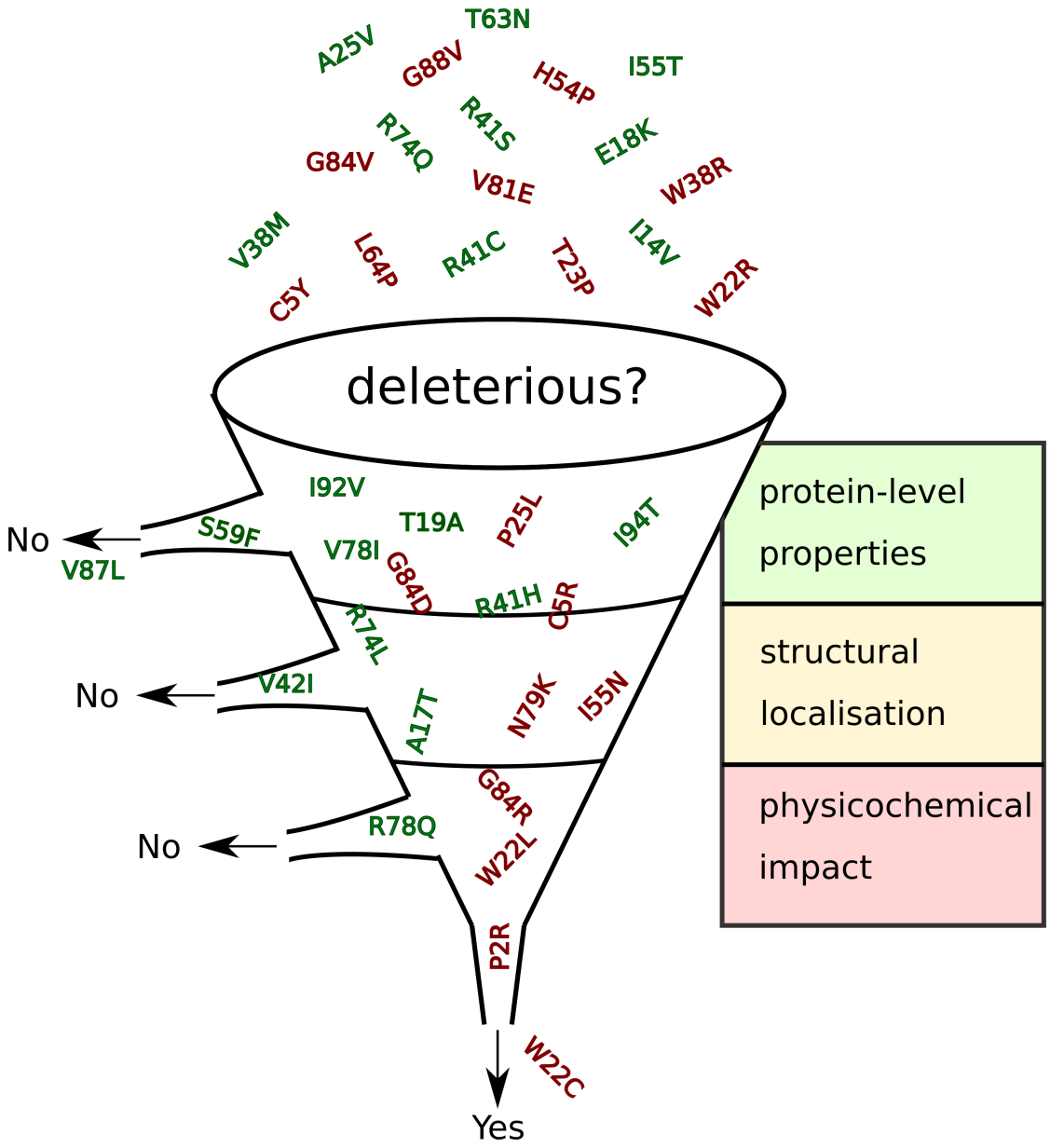

Fig. 1.4 Separating deleterious from neutral variants can be seen as a three-step process. This involves establishing whether a protein a variant localises to is likely to harbour pathogenic mutations, whether the 3D position of a variant suggests it may be disease-causing, and whether the physicochemical nature of the mutation suggests it is deleterious.

Chapter 3. A by-product of this analysis has been the creation of the ZoomVar database (http://fraternalilab.kcl.ac.uk/ZoomVar/), an online tool which allows users to annotate variants according to their localisation to protein structure and protein-protein interaction sites, in addition to the calculation of the enrichment of missense variants in different protein structural regions.

Through the work presented in Chapter 3, it becomes apparent that although features can be used to separate disease-associated from population variants, as the feature distributions for these two classes overlap, problem cases remain. For example, although disease-associated variants are clearly enriched in protein cores, and population variants depleted in these regions, a small portion of population variants are found in protein cores. This leaves us with a problem - how can we distinguish rare disease-associated variants from rare neutral variants if they have superficially similar characteristics? We explore this problem in Chapter 4, where we turn to protein dynamics, due to the promise this holds (discussed in Section 1.4.2). Initially, we sought to incorporate dynamics features from elastic network models in the prediction of variant impact, as we believed this would improve performance. As work from the Bahar lab (Ponzoni and Bahar, 2018) recently proved this to be the case, we decided to extend our work to the investigation of coarse-grained and atomistic molecular dynamics approaches. Here we focussed on whether modelling the chemical nature of the change, by calculating features which describe differences between wild-type and mutant trajectories, could improve impact prediction. To test this we used a dataset constructed of titin variants with known disease associations and population titin variants (the majority of which are rare). We believe this is the first example of work which uses features derived from molecular dynamics simulations within a machine learning-based framework for predicting the impact (deleterious/neutral) of missense variants. Moreover, we believe this approach offers much scope for future development, as discussed in Chapter 5.

Chapter 2

TITINdb

Missense variants which localise to the giant protein titin are particularly difficult to classify. As discussed in Section 1.3.2, since the advent of NGS technology a number of titin variants have been associated with both skeletal and cardiac forms of myopathy (Chauveau et al., 2014b; Hastings et al., 2016; Helle and Parikh, 2016; Savarese et al., 2016). However, due to titin's large size, even most healthy individuals possess rare or unique titin variants (Lopes et al., 2013); therefore distinguishing disease-associated variants from neutral variants is a non-trivial task. To complicate matters further, it has recently been shown that certain titin variants can be pathogenic in particular constellations or act as phenotype modifiers (Evilä et al., 2014). For example, in recent years novel paediatric forms of core myopathy with heart disease have been associated with the inheritance of a rare titin missense variant in trans with a truncating variant (Chauveau et al., 2014a). As a large number of titin variants of unknown significance exist (Campuzano et al., 2015; Lopes et al., 2013; van Spaendonck-Zwarts et al., 2014), the correct evaluation of these could contribute towards elucidating the problem of missing heritability. However, at the start of this project, the *in silico* assessment of such variants was hampered by the large size of the titin protein and discrepancies in titin domain boundaries and numbering, in addition to the incomplete structural coverage of titin (Laddach et al., 2017). To overcome these problems, and to facilitate the interpretation of titin missense variants, we created the web application TITINdb. This integrates titin structure, variant, sequence and isoform information, along with pre-computed predictions of the impact of non-synonymous single nucleotide variants. Key steps in the creation of this tool included defining titin domain

boundaries and the large scale homology modelling of titin Fn3 and Ig domains, in order to allow for the structural mapping and further *in silico* assessment of titin missense variants.

This chapter consists of the paper "TITINdb-a computational tool to assess titin's role as a disease gene" published in the journal *Bioinformatics* which describes this resource. The project was conceived by Prof. Franca Fraternali and Prof. Mathias Gautel. I carried out all the work, under their guidance, including homology modelling, analysis, creation and deployment of the web application, and writing of the paper. It should be noted that references for the paper and its supplementary materials are included separately and directly succeed each of these. All references also feature in the bibliography of this thesis.

Bioinformatics, 33(21), 2017, 3482–3485
doi: 10.1093/bioinformatics/btx424
Advance Access Publication Date: 4 July 2017
Applications Note

OXFORD

Structural bioinformatics

TITINdb—a computational tool to assess titin's role as a disease gene

Anna Laddach, Mathias Gautel and Franca Fraternali*

Randall Division of Cell and Molecular Biophysics, King's College London BHF Centre of Research Excellence, London SE1 1UL, UK

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on March 10, 2017; revised on June 15, 2017; editorial decision on June 21, 2017; accepted on July 3, 2017

Abstract

Summary: Large numbers of rare and unique titin missense variants have been discovered in both healthy and disease cohorts, thus the correct classification of variants as pathogenic or non-pathogenic has become imperative. Due to titin's large size (363 coding exons), current web applications are unable to map titin variants to domain structures. Here, we present a web application, TITINdb, which integrates titin structure, variant, sequence and isoform information, along with pre-computed predictions of the impact of non-synonymous single nucleotide variants, to facilitate the correct classification of titin variants.
Availability and implementation: TITINdb can be freely accessed at http://fraternalilab.kcl.ac.uk/TITINdb
Contact: franca.fraternali@kcl.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The giant protein titin, encoded by the gene TTN, is 35 991 amino acids in length [inferred complete (IC) isoform], weighs over 4000 kDa and spans half a sarcomere. Since the advent of next generation sequencing (NGS) technology, a number of titin missense variants, both recessive and dominant, have been associated with disease (both skeletal and cardiac forms of myopathy) (Chauveau *et al.*, 2014a; Hastings *et al.*, 2016; Helle *et al.*, 2016; Savarese *et al.*, 2016), including those which can lead to sudden cardiac death [e.g. hypertrophic cardiomyopathy (HCM)]. Unfortunately, due to titin's large size, even the majority of healthy individuals possess one or more rare titin missense variants (Lopes *et al.*, 2013). This results in the paradox that rare titin variants are commonly found; therefore, pathogenicity cannot be inferred from frequency alone. To complicate matters further, it has recently been shown that certain titin variants can be pathogenic in particular constellations or act as phenotype modifiers (Evilä *et al.*, 2014). One such scenario is the inheritance of a truncating variant along with a rare or unique missense variant in compound heterozygosity [as has been observed in childhood core myopathy with heart disease, with rare recessive mutations also found in the general population (Chauveau *et al.*, 2014b)]. In light of this information, we believe the assessment of the impact of non-synonymous single nucleotide variants (nsSNVs) at the molecular level to be essential, and propose that *in silico* analyses can be used to prioritise variants for further experimental investigation. We have created TITINdb to facilitate such prioritization.

2 Implementation and features

TITINdb includes disease-associated nsSNVs reported in the literature as well as population nsSNVs from the gnomAD database (Lek *et al.*, 2016) and 1000 genomes project (Auton *et al.*, 2015). Additionally, *in silico* saturation mutagenesis has been performed to allow users to access predictions for the impact of any possible single amino acid variants (SAVs). As experimental structures were only available for 23 of titin's 302 globular domains (132 Fn3, 169 Ig, 1 Kinase), an automated pipeline based on the Modeller software (Webb and Sali, 2016) was set up to model all 279 domains without structure (see Supplementary Figure S3 for more details). As major bioinformatics resources did not agree on titin domain numbers and boundaries, we found it necessary to define these prior to modelling.

As illustrated in Supplementary Figure S6, the TITINdb pipeline has greatly increased the structural coverage of titin domains and the quality of the coverage.

Sequence-based prediction of the impact of all nsSNVs was performed using the Condel software (González-Pérez and López-Bigas, 2011). The *in silico* assessment of the impact of nsSNVs using structural information was performed using the DUET software (Pires *et al.*, 2014a) for all known nsSNVs which map to domain structures, additionally the mCSM software (Pires *et al.*, 2014b) was used to predict the impact of all possible SAVs. Where experimental structures of titin domains in complex with binary interaction partners exist, mCSM was also used to predict the impact of SAVs on protein–protein binding affinity. Other structural analysis provided by the application includes computation of the quotient solvent accessible surface area [Q(SASA)] of all residues which map to structure [calculated using POPS (Cavallo *et al.*, 2003)] and predictions of which residues are involved in protein–protein interactions [calculated using SPIDDER (Porollo *et al.*, 2007)]. Of note, however, is the absence of experimental, molecularly resolved protein–protein interaction data for most of titin's domains, precluding detailed impact analysis on protein–protein interactions. Additionally, nsSNVs are annotated with functional site information from UniProt (The UniProt Consortium, 2017), including residue modifications.

Representative structures for each domain were used in the computation of all structural analyses, apart from the calculation of Q(SASA). This was calculated separately for each structure. A list of structure representatives can be found in the Supplementary Tables S1 and S2.

The application enables users to perform a number of visualizations, which include viewing population nsSNVs as distributions on structures, colour-coded by minor allele frequencies. Additionally, users are able to confidentially upload their own structure for nsSNV visualization (this may be useful if a group has an unpublished crystal structure or believe their own model to be of better quality).

All structures and *in silico* analyses can be freely accessed and downloaded. Additionally, we provide quality assessment of the models (in the form of zDOPE scores and per-residue DOPE plots) (Shen and Sali, 2006) along with the alignments used for homology modelling.

Video tutorials showing the use of TITINdb can be found at http://fraternalilab.kcl.ac.uk/TITINdb/tutorials/

3 Applications

3.1 Investigating disease associated nsSNVs

A potential application of TITINdb that involves investigating SNVs associated with specific diseases is shown in Figure 1 and further explored in section S2.1 of the Supplementary Materials. The facility to search by disease enables the detection of patterns or hotspots characteristic of variants associated with particular diseases. Two known nsSNV hotspots exist: one in domain Fn3-119 associated with hereditary myopathy with early respiratory failure (HMERF) (Pfeffer *et al.*, 2015) and one in Ig-169 associated with tibial muscular dystrophy, limb-girdle muscular dystrophy 2J (TMD/LGMD2J) (Chauveau *et al.*, 2014a; Hackman *et al.*, 2002; Savarese *et al.*, 2016).

TITINdb facilitates the visualization of nsSNVs associated with these diseases on structure (see Fig. 1); for both conditions nsSNVs can be observed to cluster in 3D space. In each case it can also be clearly seen that the distribution of disease associated nsSNVs on 3D structure is distinct from the distribution of population nsSNVs from the gnomAD database. Furthermore it becomes clear that all disease-associated nsSNVs discussed here are fairly buried (as indicated by a burgundy colour); therefore, it appears likely that they may disrupt protein stability. From the pre-calculated *in silico* analysis, it can be seen that all these disease associated nsSNVs are predicted to be destabilizing by DUET (Pires *et al.*, 2014a). TMD associated nsSNVs are also predicted, by mCSM (Pires *et al.*, 2014b), to disrupt the interaction between titin and obscurin, albeit by varying magnitudes; this has been validated experimentally (Fukuzawa *et al.*, 2008; Rudloff *et al.*, 2015). Interestingly the I35947N variant is predicted to have the least impact on the titin-obscurin interaction affinity (mCSM score -0.17 kcal/mol) out of all the TMD associated variants; this correlates with *in vitro* experimental observations where negligible differences have been found between this variant and wild-type titin (Fukuzawa *et al.*, 2008; Rudloff *et al.*, 2015). Additionally, the majority of HMERF associated nsSNVs are predicted to be deleterious by Condel (González-Pérez and López-Bigas, 2011), whereas only half the TMD associated nsSNVs are predicted to be deleterious. This highlights the need to take into consideration multiple sources of information, as provided by TITINdb, when predicting the potential impact of nsSNVs, and does not exclude experimental validation on a case-by-case basis.

Despite being a hotspot for HMERF associated nsSNVs, no experimental PDB structures or models are currently publicly available for the domain Fn3-119. Therefore, TITINdb has made possible the visualization of HMERF nsSNVs on structure and the *in silico* prediction of their impact at the molecular level. Multiple PDB structures exist for the domain Ig-169 (commonly referred to as M-10), to which TMD associated nsSNVs localize. Here, users can select which structure they wish to use to perform nsSNV visualization.

3.2 Investigating NGS nsSNV data

An application of TITINdb we believe to be particularly useful is the analysis of variants from NGS data. Specifically, the tool can be used in the prioritization of rare variants observed in disease cohorts for further experimental investigation. An example of such a variant is the P13979S titin N2B (isoform) nsSNV, which is published in the Supplementary Information associated with the article from Lopes *et al.* (2013), and further described in Section S2.2 of the Supplementary Materials. The nsSNV is found in 3/143 patients with HCM, leading to a cohort minor allele frequency (MAF) of between 0.01 and 0.02 (details on zygosity are not available).

A notable feature of TITINdb is the ability to search by different isoform positions. Tools such as ANNOVAR (Wang *et al.*, 2010) enable researchers to map variants from genomic to protein coordinates, however, depending on the protocol followed, variants may be mapped to different isoforms. For our nsSNV of interest, the N2B isoform coordinate is reported, thus the 'search by position' facility allows it to be mapped to both other major isoforms and the position within the affected domain. Additionally, it can be seen that the nsSNV localizes to residue position 5 of domain Fn3-55 and is present in all isoforms apart from the novex-3 isoform; therefore it is expressed in both cardiac and skeletal muscle.

TITINdb allows easy access to information concerning the nsSNV's potential impact. Structurally, it can be seen that the affected residue has a Q(SASA) of 0.1 (this information is provided in the table on the nsSNV page), indicating that it is buried and that the nsSNV could potentially cause disease through destabilization of

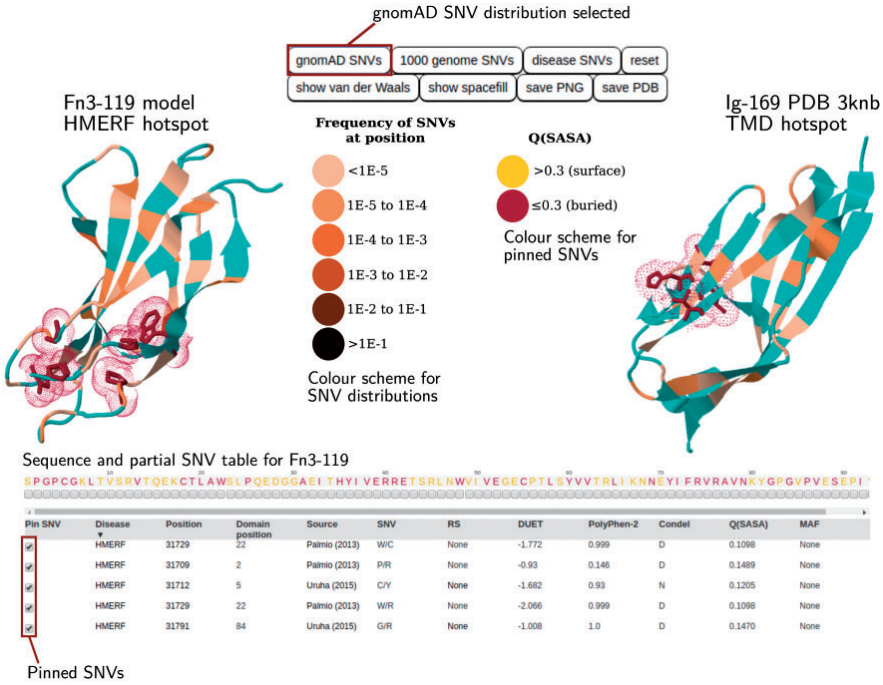

Fig. 1. TITINdb user interface overview. The HMERF and TMD associated nsSNV hotspots are shown. Users can pin disease-associated nsSNVs from the SNV table onto domain structure and visualize these against the distribution of population nsSNVs (Gnomad or 1000 genomes). Pre-computed *in silico* analyses are shown in the SNV table (more information can be accessed by scrolling horizontally and vertically)

the underlying domain. It can also be seen, from predictions by the software SPPIDER (Porollo *et al.*, 2007), that the affected residue is not predicted to be involved in protein-protein interactions and thus is unlikely to cause disease through the disruption of these.

On comparison to known nsSNVs it can be seen that the nsSNV is present in both the 1000 genomes data and the gnomAD database with MAFs of 3.7036E-03 and 9.98403E-04. This indicates the variant is rare but present in a small proportion of nominally healthy individuals. From the MAF values it can be deduced that the variant is enriched in the HCM cohort (which we know has a MAF between 0.01 and 0.02). This suggests that the variant is either neutral, disease-causing with incomplete penetrance, recessive, or that a small number of nominally healthy individuals have undiagnosed HCM.

Structure (DUET)- and sequence (Condel)-based predictions of the impact of the nsSNV can be observed. The DUET score of −2.703 kcal/mol suggests the variant is highly destabilizing and supports the hypothesis derived earlier from the Q(SASA) that the variant could potentially lead to disease by disrupting the domain structure. Furthermore, it can be seen that the variant is also predicted to be deleterious by Condel.

As no experimental structures exist for the domain Fn3-55, 3D visualization and access to pre-computed structural analyses are made possible by the homology model provided as part of TITINdb. One salient feature is that, if nsSNVs are pinned on structure from the sequence, any related/identical nsSNVs rise to the top of the nsSNV table and become highlighted in either yellow (surface) or red (buried) according to their Q(SASA) (see Supplementary Figure S2); the pinned nsSNVs also follow this colour scheme.

The results indicate that, although the analysed nsSNV is highly likely to affect the domain structure, it is unclear whether this will contribute to the disease phenotype (primarily as the mutant titin may not be expressed *in vivo* in heterozygous cases).

Further information concerning applications of TITINdb can be found in the Supplementary Materials. In particular, it is hoped that the tool will enable clinicians to perform the information-based assessment of variants from patient data, and assist biologists in the prioritization of domain structures for biophysical characterization.

Funding

This work has been supported by the British Heart Foundation [RE/13/2/30182, RG/15/8/31480 and CH/08/001] and the Biotechnology and Biological Sciences Research Council [BB/H018409/1 to FF].

Conflict of Interest: none declared.

References

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Cavallo,L. *et al.* (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.*, **31**, 3364–3366.

Chauveau,C. *et al.* (2014a) A rising titan: TTN review and mutation update. *Hum. Mutat.*, **35**, 1046–1059.

Chauveau,C. *et al.* (2014b) Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum. Mol. Genet.*, **23**, 980–991.

Evilä,A. *et al.* (2014) Atypical phenotypes in titinopathies explained by second titin mutations. *Ann. Neurol.*, **75**, 230–240.

Fukuzawa,A. *et al.* (2008) Interactions with titin and myomesin target obscurin and obscurin-like 1 to the M-band: implications for hereditary myopathies. *J. Cell. Sci.*, **121**, 1841–1851.

González-Pérez,A. and López-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.

Hackman,P. *et al.* (2002) Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. *Am. J. Hum. Genet.*, **71**, 492–500.

Hastings,R. *et al.* (2016) Combination of whole genome sequencing, linkage, and functional studies implicates a missense mutation in titin as a cause of autosomal dominant cardiomyopathy with features of left ventricular noncompaction. *Circ. Cardiovasc. Genet.*, **9**, 426–435.

Helle,E. *et al.* (2016) Wrestling the giant: new approaches for assessing titin variant pathogenicity. *Circ. Cardiovasc. Genet.*, **9**, 392–394.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.

Lopes,L. *et al.* (2013) Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *J. Med. Genet.*, **50**, 228–239.

Pfeffer,G. *et al.* (2015) Diagnosis of muscle diseases presenting with early respiratory failure. *J. Neurol.*, **262**, 1101–1114.

Pires,D. *et al.* (2014a) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*, **42**(Web Server issue), W314–W319.

Pires,D. *et al.* (2014b) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.

Porollo,A. *et al.* (2007) Prediction-based fingerprints of protein-protein interactions. *Proteins*, **66**, 630–645.

Rudloff,M.W. *et al.* (2015) Biophysical characterization of naturally occurring titin M10 mutations. *Protein Sci.*, **24**, 946–955.

Savarese,M. *et al.* (2016) Increasing role of titin mutations in neuromuscular disorders. *J. Neuromuscul. Dis.*, **3**, 293–308.

Shen,M. and Sali,A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, **15**, 2507–2524.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.*, **86**, 2.9.1–2.9.37.

Supplementary Materials

S1 Application features

TITINdb enables users to:

- access pre-computed predictions of the impact of gnomAD, 1000 genomes and disease-associated nsSNVs calculated using Duet (Pires et al., 2014b) and Condel (González-Pérez and López-Bigas, 2011)
- access pre-computed predictions of the impact of any possible single amino acid variants (SAVs) which localise to titin domain structures, calculated using mCSM (Pires et al., 2014a), and the impact of any possible nsSNVs calculated using Condel (González-Pérez and López-Bigas, 2011)
- access pre-computed predictions of the impact of any possible SAVs on protein-protein binding affinity calculated using mCSM (Pires et al., 2014a) (where experimental structures for titin domains in complex with interaction partners exist)
- visualise nsSNVs on structure and save images as PNG files
- visualise 3D nsSNV distributions on structure from the 1000 genomes project, gnomAD and disease-associated nsSNVs
- download all nsSNV information for nsSNVs which localise to a titin Ig, Fn3 or kinase domains as a CSV file
- explore disease hotspots through the facility to "search by disease"
- access structural analysis, i.e. the Q(SASA) for each residue (computed using POPS (Cavallo et al., 2003)) and predictions of which residues take part in protein-protein interactions (computed using SPPIDER (Porollo et al., 2007))
- access Uniprot (The UniProt Consortium, 2017) functional site annotations, including residue modifications
- download structures and models of titin Ig, Fn3 and kinase domains as PDB files
- assess model quality by using zDOPE score, per residue DOPE score plots and query-template alignments
- upload structures/models for nsSNV visualisation
- translate between isoform amino acid positions for all seven titin isoforms with RefSeq sequences (IC, N2AB, N2A, N2B, novex-1, novex-2, novex-3)

S2 Case studies

Mutations which result in amino acid changes can be broadly divided into two classes: those which are present in the population and which are generally not phenotypically deleterious, or those which are causative of disease. Certain mutations may also be unclassified, and of uncertain effect. For titin this is frequently the case when nsSNVs are observed in disease cohorts but a causative link with the condition has not been established. In TITINdb we present analyses of nsSNVs which fall into the "classified" classes, to enable users to better understand the properties of such nsSNVs, and in order to facilitate the classification of currently unclassified nsSNVs; a graphical summary of this can be seen in Fig. S1. The two classes are, however, not mutually exclusive. Some disease-associated nsSNVs may demonstrate incomplete penetrance or may be recessive (e.g. in compound heterozygosity (Evilä et al., 2014; Chauveau et al., 2014a)), or play a modifier role in disease.

Conversely, it is possible that certain nsSNVs may be misclassified due to nominally healthy individuals harbouring undiagnosed disease, or linkage disequilibrium existing between variants which are actually disease-causing and those which are not. Indeed, it must be remembered that the gnomAD database, although not expected to be enriched in disease-associated nsSNVs, does contain genetic information from disease cohorts, although individuals with severe paediatric disease have been filtered out (Lek et al., 2016). Nevertheless, the gnomAD database is expected to contain rare recessive disease-causing TTN mutations similar to any gene, e.g. recessive autosomal CFTR mutations causing cystic fibrosis (frequency of heterozygous carriers between 1% and 3.5% depending on population (Strom et al., 2011)) or recessive X-linked DMD mutations causing Duchenne muscular dystrophy (carrier frequency 0.18% Mundy et al. (2016)). The main difference is that such recessive mutations are expected to be more numerous in titin due to the large size of the coding sequence (>100kb). *In silico* saturation mutagenesis has also been carried out to enable users to predict the impact of novel nsSNVs.

In the following case studies, we show how TITINdb allows the exploration of disease associated nsSNVs, and how the application can be used to facilitate the classification of new nsSNVs.

S2.1 Investigating disease associated nsSNVs

TITINdb allows the user to search by disease; currently 12 myopathies have associated titin variants. This enables the detection of patterns or hotspots characteristic of variants associated with particular diseases. Two known nsSNV hotspots have been investigated in the main text: one in domain Fn3-119 associated with hereditary myopathy with early respiratory failure (HMERF) (Palmio et al., 2014; Pfeffer et al., 2012; Ohlsson et al., 2012; Toro et al., 2013; Izumi et al., 2013; Vasli et al., 2012; Uruha et al., 2015) and one in Ig-169 associated with tibial muscular dystrophy/limb-girdle muscular dystrophy 2J (TMD/LGMD2J) (Pollazzon et al., 2010; Van den Bergh et al., 2003; Hackman et al., 2002; Evila et al., 2016). Here, we present an additional example, which investigates nsSNVs associated with dilated cardiomyopathy (DCM). This disease is primarily associated with titin truncating variants (Schafer et al., 2016). There are, however, a number of nsSNVs which are also associated with this condition.

On searching for dilated cardiomyopathy (DCM) associated nsSNVs (Itoh-Satoh et al., 2002; Gerull et al., 2002; Herman et al., 2012; LIU et al., 2008; Roncarati et al., 2013; Matsumoto et al., 2005), it can be seen that these do not appear to form a distinct hotspot as observed for both TMD and HMERF, and, although two of the thirteen nsSNVs localise to the domain Ig-30, they are situated at opposite ends of the domain. It can be noted that these nsSNVs are both found in gnomAD and one is also found in the 1000 genomes project; indicating that these nsSNVs are unlikely to demonstrate complete penetrance. Upon visualisation on the available model of Ig-30, it can be seen that the affected residues are also both located on the surface of the protein, with the S4780N nsSNV not predicted to be destabilising. This suggests that pathogenicity could be a result of the disruption of protein-protein interactions. Indeed, from the SNV table it can be seen that both residues affected by these nsSNVs are predicted to take part in protein-protein interactions.

From this example and those in the main text, it can be seen that inferring titin nsSNV pathogenicity can be a complex task; however, even with the sparse association of titin nsSNVs with disease, emerging patterns/hotspots are already discernible and further work will aid in the classification of currently unclassified nsSNVs.

S2.2 Investigating NGS nsSNV data

Hypertrophic cardiomyopathy (HCM) is a clinically heterogeneous disease in which the walls of the heart become thickened, in particular the wall of the left ventricle (Pantazis et al., 2015). It is caused by cardiac sarcomere mutations with incomplete penetrance (Seidman and Seidman, 2011) and is one of the leading causes of sudden death in young adults, affecting approximately 1 in 500 individuals (Maron et al., 1995). A number of rare and unique missense variants have recently been found in a cohort of HCM patients (Lopes et al., 2013), however how many of these play a causative

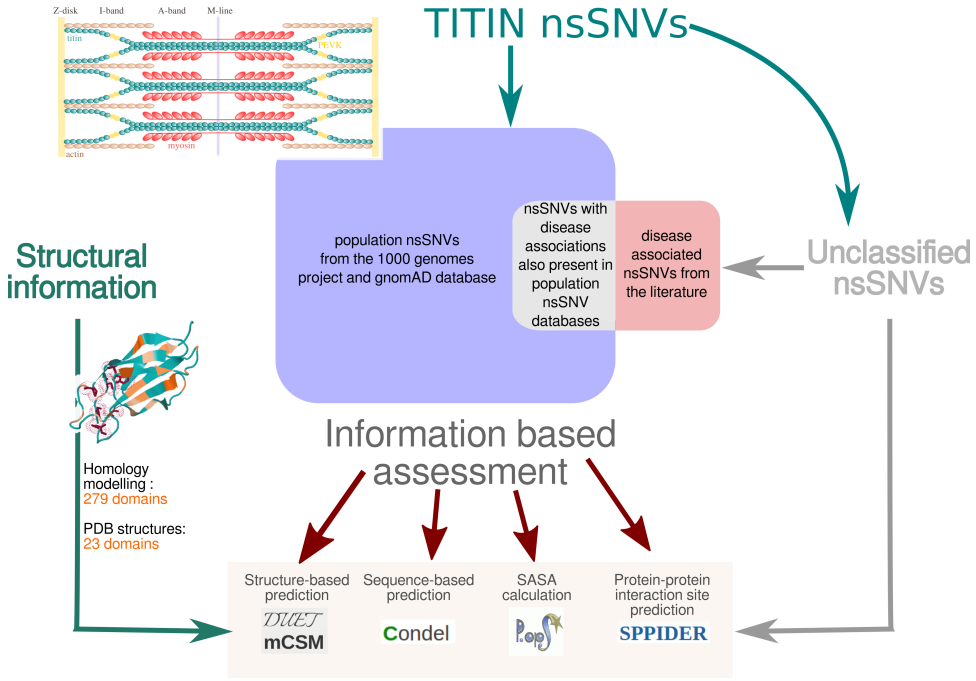

Figure S1: Summary showing how the information in TITINdb can be used to assess unclassified nsSNVs. This can be done in several ways, for example by comparing nsSNV properties to known SNVs, of which there are currently 51 disease-associated nsSNVs and a larger number of population nsSNVs (19741 in the gnomAD database and 1982 in the 1000 genomes data). There is an overlap of 24 nsSNVs between disease-associated nsSNVs and population nsSNVs (from the 1000 genomes project and gnomAD database); this is most likely due to the incomplete penetrance of some disease-associated SNVs. Information based assessment can also contribute to the classification of unclassified SNVs through structure- and sequence-based predictions of their impact, and properties of their location on 3D structure.

role in the disease is currently unclear. In the main text we show one possible use of TITINdb through the analysis of the P13979S titin N2B nsSNV which is published in the supplementary information associated with the paper from Lopes et al. (2013). The visualisation of this particular nsSNV mapped to structure can be seen in Fig. S2.

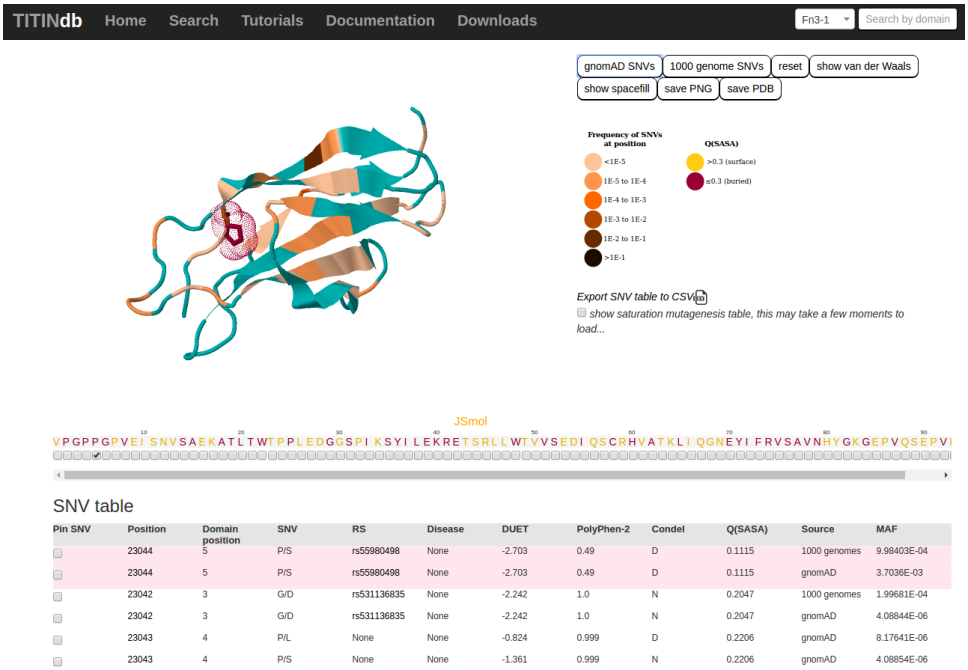

Figure S2: Visualisation of the P13979S titin N2B nsSNV on the Fn3-55 domain structure. The red colour indicates the nsSNV affects a buried residue. The nsSNV can be compared to the gnomAD distribution of nsSNVs (shown in orange).

S3 Methods

S3.1 Data

1000 genomes titin variant data was obtained in VCF format using the online data slicer (Auton et al., 2015). Variants were mapped to isoform and position using ANNOVAR (Wang et al., 2010). gnomAD variant data was obtained from the gnomAD web server (Lek et al., 2016) in Variant Dataset Format and titin nsSNVs were extracted using Hail (Ganna et al., 2016). Titin related disease nsSNVs were obtained from 'A rising titan: TTN review and mutation update' (Chauveau et al., 2014b). Disease nsSNVs reported in the literature discovered since the publication of Chauveau et al. (2014b) were queried for on PubMed using the terms "("titin"[All Fields]) AND ("snp"[All fields])","("titin"[All Fields]) AND ("mutation"[All fields])" and "("titin"[All Fields]) AND ("variant"[All fields])".

Data from the 1000 genomes project originates from 2504 nominally healthy individuals. The gnomAD database represents a much larger number of individuals (138632) and is aggregated from a number of studies, including the 1000 genomes project. It does not only include healthy individuals; however individuals with severe paediatric diseases have been filtered from the dataset. Therefore the gnomAD database is unlikely to be enriched in variants associated with severe diseases and is likely to be indicative of a population distribution of variants.

Titin functional site information was obtained from UniProt (The UniProt Consortium, 2017).

S3.2 Defining titin domain boundaries

HMMER (Finn et al., 2011) was used to scan the protein sequence of titin IC variant (NP_001254479.2), obtained from the RefSeq database (Pruitt et al., 2012) against Pfam seed libraries (Finn et al., 2014). Where hits overlapped the hit with the lowest E-value was accepted. When the lowest E-value hit for a region was greater than 0.0001 additional evidence was required to accept a hit, such as an existing experimental structure or homology to other titin domains of the same type. This homology was assessed by creating an HMM from all titin domains of the same type (which had been identified with an E-value <0.0001) and rescanning the titin protein sequence against this HMM. If identified domains were detected with an E-value <0.0001, upon scanning against this titin-specific HMM, they were considered homologous.

Sequences of titin domains defined in this way, including an extra 5 amino acids upstream and 16 amino acids downstream of the Pfam defined boundary, were aligned using T-coffee (Notredame et al., 2000) (separate alignments were created for titin Fn3 and Ig domain sequences).

Sequence logos were also created from these alignments using Weblogo (Notredame et al., 2000). Additionally, sequence logos were created from the Pfam I-set and Fn3 seed alignments. It should be noted that all titin Ig domains are determined to be of the I-set type when defined by scanning against Pfam seed alignments.

Comparisons were made between sequence logos derived from aligned titin domain sequences and those derived from Pfam seed alignments. Where the two types of sequence logo appeared to differ substantially or the domain boundaries did not appear to be clearly defined by sequence conservation, the boundaries were mapped onto available experimental 3D structures and information from this mapping used to redefine titin domain boundaries. It must be noted that differences between sequence logos, and clarity of the initial domain boundaries, were assessed heuristically by eye. Additionally, the redefinition of domain boundaries was accomplished heuristically by setting these visually to cover the entire globular portion of the mapped domain structures.

S3.3 Mapping of titin isoforms

Stretcher (Myers and Miller, 1988) was used to align all titin isoforms to the IC variant. Isoform sequences were obtained from RefSeq (Pruitt et al., 2012). Positions were mapped according to these alignments.

S3.4 Modelling of titin domains

An automated homology modelling pipeline was set up. The pipeline takes a fasta file of domain sequences as input and uses only publicly available PDB structures as templates. The overall modelling process can be seen in Fig. S3A and a flow diagram detailing the template selection process is depicted in Fig. S3B. The template search, modelling and model assessment were performed using Modeller (Webb and Sali, 2016), the alignment of query and templates performed using 3DCoffee (O'Sullivan et al., 2004), and the overall pipeline produced using Python 2.7. Models were selected based on zDOPE score. zDOPE score is a normalised atomic distance-dependent statistical potential based on a sample of native structures (Shen and Sali, 2006). Lower zDOPE scores indicate better models with zDOPE scores below -1 indicating the distribution of atomic distances is similar to that in the sample of native structures.

The I-TASSER server (Zhang, 2008) was used to model Ig-112 as a satisfactory (negative) zDOPE score was not obtained using the homology modelling pipeline.

S3.5 Validation of models

The modelling pipeline described in Section S3.3 was used to model all models with existing experimental structures. However so as to exclude the already solved structure from the templates, all hits with an identity >95% were excluded during template selection.

To validate the models these were compared to the representative experimental structures detailed in Table S1 in a similar manner to Sánchez and Sali (1998) as well as the Critical Assessment of Protein Structure Prediction experiments (Cozzetto et al.,

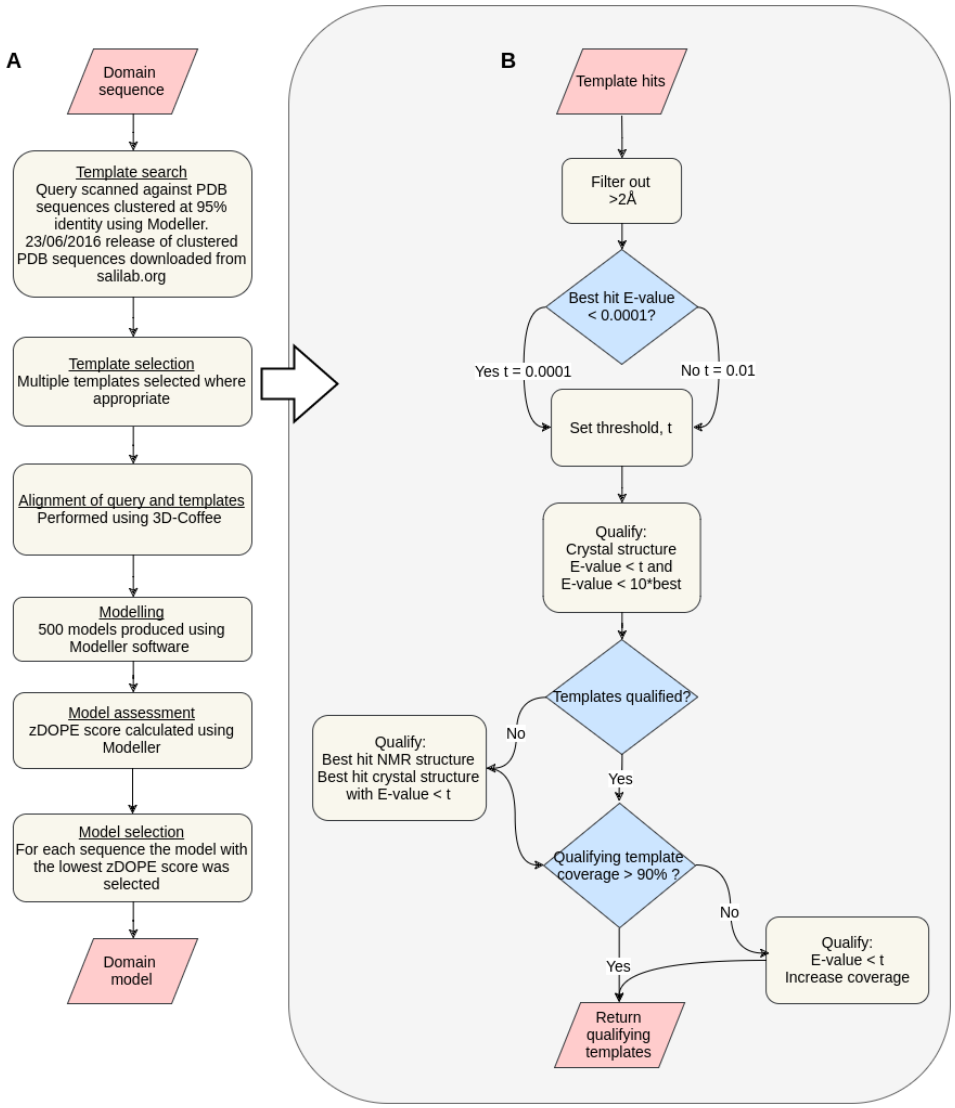

Figure S3: Flow-diagrams showing **A** an outline of the modelling pipeline and **B** template selection criteria. Qualifying templates refer to those templates which are selected for the modelling procedure.

2009). Sequence alignments between structures and models were calculated using Muscle (Edgar, 2004) and, structural superpositions guided by the sequence alignments as well as per residue

Table S1: PDB structures used for the structural investigation of nsSNVs.

domain	PDB ID	method	resolution/ Å
Ig-1	2a38	X-RAY	2.00
Ig-2	2a38	X-RAY	2.00
Ig-10	1g1c	X-RAY	2.10
Ig-18	5jdd	X-RAY	1.53
Ig-19	5jdd	X-RAY	1.53
Ig-20	5jdd	X-RAY	1.53
Ig-84	5j0e	X-RAY	2.00
Ig-94	1waa	X-RAY	1.80
Ig-156	3lcy	X-RAY	2.50
Ig-157	3lcy	X-RAY	2.50
Ig-158	2j8h	X-RAY	1.99
Ig-159	2j8h	X-RAY	1.99
Ig-160	2bk8	X-RAY	1.69
Ig-163	3qp3	X-RAY	2.00
Ig-164	1tnn	NMR	NA
Ig-166	3puc	X-RAY	0.96
Ig-169	3knb	X-RAY	1.40
Fn3-3	4o00	X-RAY	1.85
Fn3-62	1bpv	NMR	NA
Fn3-66	3lpw	X-RAY	1.65
Fn3-67	3lpw	X-RAY	1.65
Fn3-132	2nzi	X-RAY	2.90‘

RMSD calculations were performed using Theseus (Theobald and Wuttke, 2006).

S3.6 *In silico* assessment of the impact of nsSNVs

The *in silico* assessment of known nsSNVs occurring within Ig and Fn3 domains was performed using DUET (Pires et al., 2014b). This tool exploits structural information and is based on a consensus of mCSM (Pires et al., 2014a), a graph-based method combined with machine learning to predict free energy changes resulting from single point mutations, and SDM (Topham et al., 1997) which uses environment-specific substitution tables. The prediction of impact for all possible SAVs which localise to domain structures was carried out using mCSM (Pires et al., 2014a); this algorithm was chosen as its speed enables the prediction to be carried out for the large number of such possible SAVs (492271). Where experimental structures were available these were used for the assessment. Where no experimental structures were available the model with the lowest zDOPE score was used. See Table S1 for experimental structures used in the *in silico* assessment of nsSNVs. The impact of nsSNVs on protein-protein binding affinity was also predicted using mCSM (Pires et al., 2014a) where experimental structures of titin domains in complex with binary interaction partners exist. See Table S2 for structures used in the assessment of the impact of nsSNVs on protein-protein binding affinity.

Assessment of all nsSNVs was performed using the method Condel (González-Pérez and López-Bigas, 2011). This method is purely sequence-based and uses the weighted average of the normalised scores of 5 methods: Log R Pfam E-value, MAPP, Mutation Assessor, Polyphen2 and Sift (González-Pérez and López-Bigas, 2011).

S3.7 Definition of structural elements

Interface and core regions were defined using POPS (Cavallo et al., 2003). Residues with a Q(SASA) (quotient solvent accessible surface area) > 0.3 were defined as being surface residues and those with a Q(SASA) ≤ 0.3 defined as core residues. Here Q(SASA) is defined as the quotient of the SASA (solvent accessible surface area) and Surf (surface area of the isolated residue).

Putative PPI interface regions were predicted using SPPIDER II (Porollo et al., 2007) with a balanced trade-off between sensitivity and specificity (SPPIDER estimates this based on a control data set of 149 protein chains with no sequence homology), using representative structures (see Table S1). SPPIDER II predicts interface residues with 74.18% accuracy, 60.30% recall, 63.72% precision, and a Matthews correlation coefficient (MCC) of 0.42.

S3.8 Creation of a titin database

A database was set up to integrate titin structural, variant and disease information. The database was created using SQLite and DJANGO. nsSNVs from the 1000 genomes project, ExAC and the paper ‘A rising Titin: TTN review and mutation update’ (Chauveau et al., 2014b), as well as information

Table S2: PDB structures used for the investigation of the impact of nsSNVs on protein-protein binding affinity.

domain	PDB ID	method	resolution / Å	interaction partner
Ig-1	1ya5	X-RAY	2.44	TCAP
Ig-2	1ya5	X-RAY	2.44	TCAP
Ig-169	3knb	X-RAY	1.40	OBSL1
Ig-2	4c4k	X-RAY	1.95	TCAP

concerning structures modelled by the pipeline and PDB structures were loaded to the database.

S3.9 Web server implementation

TITINdb is a python-based application implemented using the DJANGO framework. JSmol is used to visualise protein structures. It is hosted on an Apache2.2.15 web server.

S4 Results

S4.1 Comparison of the calculated properties of categorised nsSNVs/SAVs

A comparison of the calculated properties of categorised nsSNVs/SAVs which feature in TITINdb can be seen in Fig. S4. The top left plot compares the log-transformed distribution of gnomAD frequencies for nsSNVs found in the 1000 genomes project with the log-transformed distribution of gnomAD frequencies for nsSNVs associated with disease (a small pseudocount of 3.6E-6 has been added to each frequency to allow for log transformation). It can be seen that although the two subsets are significantly different the MAF values show a large overlap. Here the caveat must be taken that data from 1851 individuals from the 1000 genomes project is a subset of the gnomAD data set. However, as data from the 1000 genomes project only accounts for 1.3% of the data in gnomAD it is considered unlikely that this will have a large impact on the MAF values observed in gnomAD. These results suggest that not all pathogenic variants can be distinguished from non-disease causing variants by differences in MAF.

The top right plot shows the distributions of Condel scores for nsSNVs from the 1000 genomes project, the gnomAD database, saturation mutagenesis, and disease-associated nsSNVs. The disease-associated nsSNVs are predicted to be significantly more deleterious than nsSNVs the other subsets. nsSNVs from saturation mutagenesis are predicted to be significantly more deleterious than those from the 1000 genomes project and gnomAD database (from Fig. S4 it is clear the difference is very small and significance reached due to the large size of the dataset), however significantly less deleterious than disease-associated variants.

In the bottom left plot it can be seen that disease-associated nsSNVs localise to residues with a significantly lower Q(SASA) than nsSNVs and SAVs from all other subsets. SAVs from saturation mutagenesis localise to residues with a significantly lower Q(SASA) than variants from the 1000 genomes data or ExAC database, however it is clear that the difference is very small and significance reached due to the large size of the dataset.

The bottom right plot shows the distribution of mCSM scores (predicted impact on stability in kcal/mol, negative values indicate destabilisation) for variants from each subset of the data. Here it can be seen that disease-associated nsSNVs are predicted to be significantly more destabilising than nsSNVs from the 1000 genomes project, the gnomAD database and SAVs from saturation mutagenesis.

Please note that saturation mutagenesis has explored all variants (SAVs) which can be achieved by a single amino acid change through mCSM and POPs, however only those variants (nsSNVs) which can be achieved by a single nucleotide change through Condel.

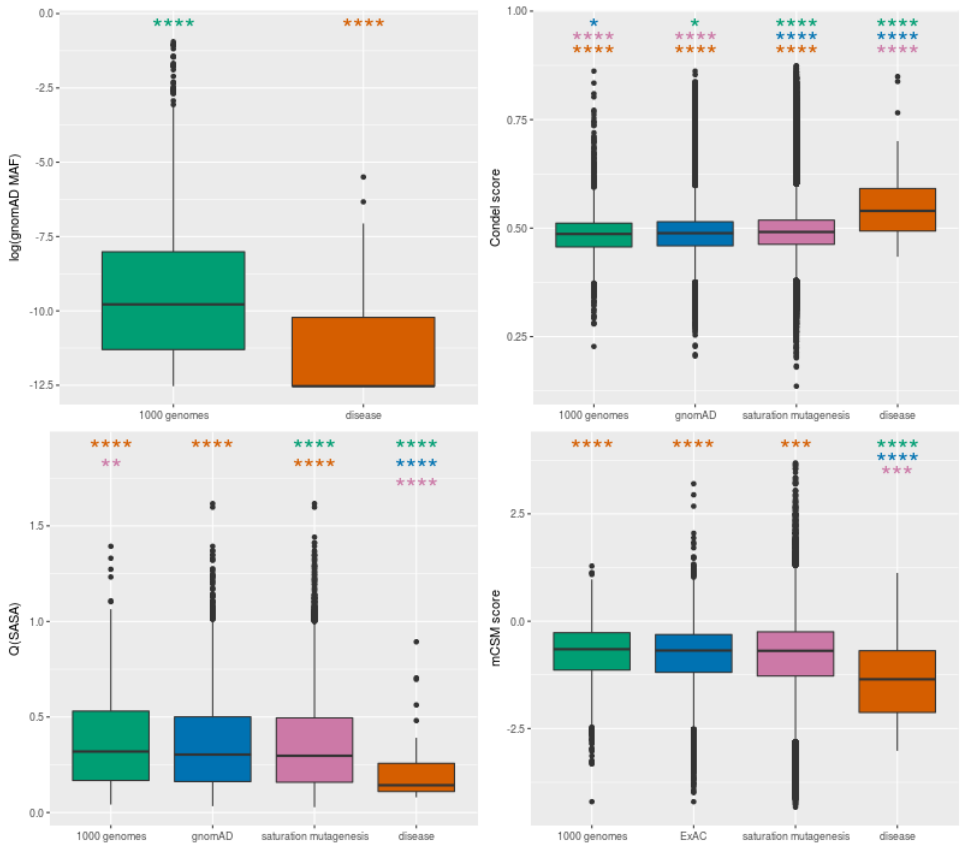

Figure S4: A comparison of the properties of titin nsSNVs from the 1000 genomes project and gnomAD database with disease-associated nsSNVs and nsSNVs/SAVs from saturation mutagenesis. It can be seen that disease-associated nsSNVs are predicted to be both significantly more destabilising and significantly more deleterious than nsSNVs/SAVs from the other subsets. It can also be seen that disease-associated SNVs localise to residues with a significantly lower Q(SASA). Significance calculated using pairwise Mann-Whitney tests with Bonferroni correction (*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001).

S4.2 Titin domain boundaries

We find titin has 169 Ig domains in contrast to the previously reported 152 (Chauveau et al., 2014b). This discrepancy arises from both the previous use of shorter isoforms to define domains rather than the reference IC isoform, and the failure to integrate available experimental evidence. Domain Ig-94 is identified with an E-value of 0.0079 which is higher than the set threshold of 0.0001, but is validated by the experimental PDB structures (1WAA, 1TIU and 1TIT) which exist of this domain. Ig-88, Ig-89, Ig-90 and Ig-98 were also identified with E-values higher than the threshold (0.00079, 0.00026, 0.0022 and 0.0005); however when the titin sequence was scanned using an HMM created from an alignment of all (165) other

titin Ig domains, these domains were identified with significantly low E-values (5.9E-11, 1.2E-12, 2.9E-10, 7.4E-12). The mapping of all titin isoform protein sequence positions to the reference IC isoform has enabled the instigation of a consistent naming scheme for domains in which Ig domains are sequentially numbered from 1 to 169 and Fn3 domains numbered sequentially from 1 to 132.

Sequences logos (see Fig. S5) showing aligned titin Fn3 sequences, differ substantially from such logos depicting Pfam seed alignments (as assessed heuristically by eye); particularly towards the end of the sequence where the conservation drops of gradually. Therefore the correct boundaries do not appear to be clearly defined from sequence alone. When mapped onto available structures, for example in the case of the Fn3 dimer 3LPW (bottom Fig. S5), it becomes clear that the Pfam defined boundaries do not cover entire titin Fn3 domains. Due to this information, it was decided the Pfam defined Fn3 domain boundaries were not accurately determined. Therefore Fn3 domains were initially identified using Pfam/HMMER and the sequences of these domains, including an extra 5 amino acids upstream and 16 amino acids downstream of the Pfam defined boundaries, were aligned using T-coffee. This alignment was cut heuristically (see Section S3.2) using structural information from available titin Fn3 crystal structures, in particular 3LFPW. An HMM was created from this alignment and titin scanned again using this HMM to redefine titin Fn3 domain boundaries.

S4.3 Modelling of titin Ig and Fn3 domains

A pipeline was set up as described in the methods section to perform the homology modelling of titin Fn3 and Ig domains. Fig. S6 shows the structural coverage of titin by experimental crystal/NMR structures from the PDB, existing models from the ModBase database (Pieper et al., 2009), and models produced by the TITINdb pipeline.

It can be seen in Fig. S6 that our pipeline has greatly increased both the structural coverage of domains and the quality of the coverage (lower zDOPE (Shen and Sali, 2006) scores indicate better structures with native structures expected to have a zDOPE score around -1). Here, for each domain sequentially along the length of titin, existing experimental structures are represented by purple diamonds and blue hexagons, ModBase models are represented by blue bars and models created by the TITINdb pipeline are represented by red bars. The closest identity each domain shares with an experimental PDB structure is annotated along the top x-axis.

Model validation was performed by modelling all domains for which experimental structures already exist, however excluding structures with >95% identity from template selection. The models were then compared to solved structures for the relevant domain as described in the methods section, in a similar manner to Sánchez and Sali (1998) and Cozzetto et al. (2009). Cumulative distribution plots for the RMSD (root-mean-square deviation) for the comparison between models and representative structures are shown in Fig. S7A (Ig domains) and Fig. S7C (Fn3 domains). It can be seen that a large proportion of modelled residues have RMSD values lower than 1Å indicating that the modelling has predicted the actual structure with a high degree of accuracy. For 82% of the models >60% of their residues fall within 1Å of the solved structure and for 65% of the models >70% of their residues fall within 1Å of the solved structure (see Table S4.3). As experimental structures have only been solved for 5 titin Fn3 domains, summary statistics are perhaps less informative, however for 80% of these >70% of residues lie within 1Å of the solved crystal structures. For this small sample size, the Fn3 models generally show less deviation in terms of alpha carbon (Cα) RMSD from the solved structures than the Ig models do. This is perhaps a result of the proportionally lower loop content of the majority of titin Fn3 domains when compared to titin Ig domains; as loop regions tend to be more flexible and less conserved these tend to result in larger RMSD values.

In the majority of cases, those residues with the highest RMSDs when compared to solved structures localise to loops; as discussed this is unsurprising due to the inherently greater flexibility of loop regions. The exception to this is the model for Ig-19, shown aligned to the crystal structure 5JDD in Fig. S7A-I, which, from the cumulative RMSD distribution appears to have been modelled with the least success. On closer inspection, it becomes apparent that the majority

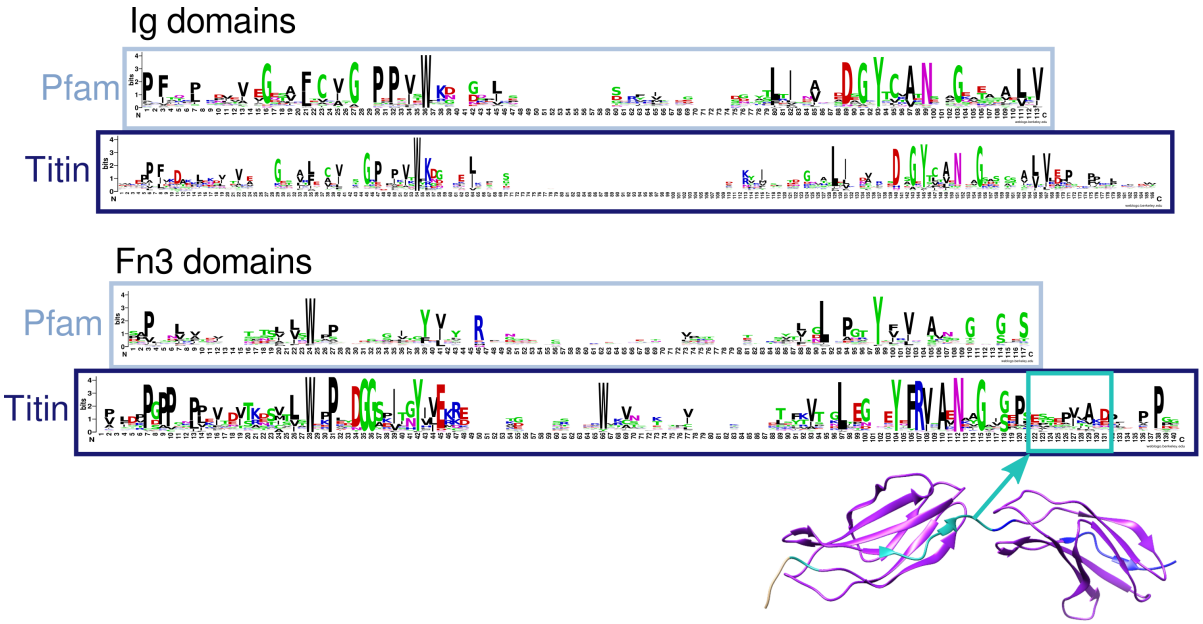

Figure S5: Sequence logos showing aligned titin domains and Pfam seed alignments for Ig and Fn3 domains. The Pfam Fn3 domain definition can be seen mapped onto the structure 3LPW in purple with structure absent from the Pfam definition in turquoise and blue.

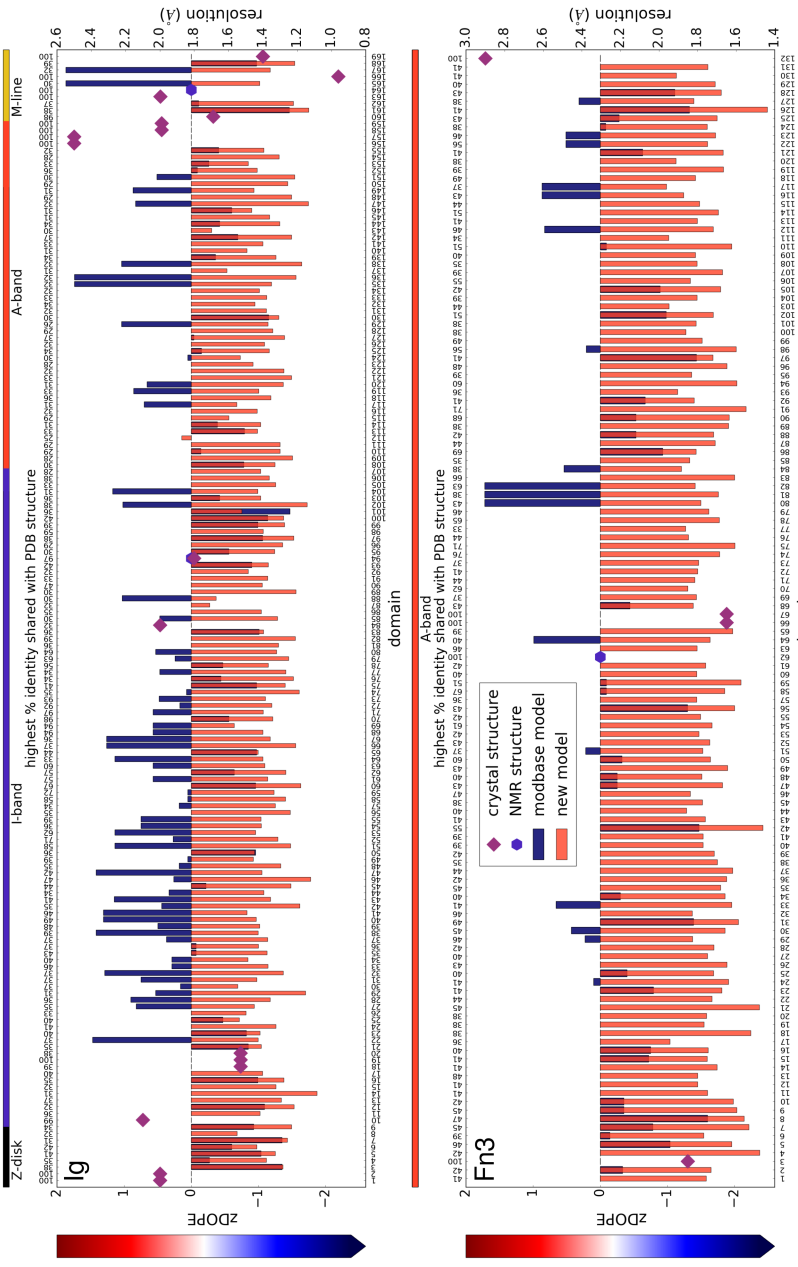

Figure S6: Structural coverage of titin Ig and Fn3 domains by models, crystal and NMR structures. New models refer to those produced during this project. Lower zDOPE scores indicate better model quality. The highest percentage identity shared with a homologous crystal structure (at the time of modelling) can be seen along the top x-axis for each domain.

of the backbone aligns closely to the crystal structure however the loop prior to the C-terminus is not as tight as in the crystal structure which results in misplacement of the final β sheet. On observation of the alignment between the query and template sequences (see Fig. S8) it can be seen that the quality of the alignment drops for the portion of the sequence which corresponds to the C-terminal β sheet which perhaps explains the poor accuracy of modelling this region. The other immunoglobulin model which appears less accurate judging by the RMSD values is Ig-164, shown aligned to the NMR structure 1TNN in Fig. S7A-II. Here, as expected, the beta sheet core regions show close alignment to the solved NMR structure 1TNN, however the loop regions show large differences, in particular a short helical stretch can be seen in the loop between beta sheets E and F in the model which is not present in the structure. Interestingly, the model was built using only crystal structures as templates, and all the crystal structures available for titin Ig domains show such a short helical segment in their analogous loops; however none of the NMR structures demonstrate this property. Therefore it is likely that the helical structure does not form in solution due to competition with the surrounding solvent for hydrogen bond formation with partially exposed residues. This indicates that the higher RMSD values observed in the loop regions for this domain may be an indicator of conformational differences caused by the distinct environments in which the structures have been solved, rather than poor model quality.

Table S3: Percentage of domains for which models have a particular percentage of residues within 1Å of the solved experimental structure after structural superposition. The analysis was performed using 17 domains for Ig domains and 5 domains for Fn3 domains (see Table S1 for domains and experimental structures used for validation).

% residues <1Å RMSD from representative structure	% Ig domains
10	100
20	100
30	100
40	100
50	94
60	82
70	65
80	35
90	12

% residues <1Å RMSD from representative structure	% Fn3 domains
10	100
20	100
30	100
40	100
50	100
60	80
70	80
80	60
90	0

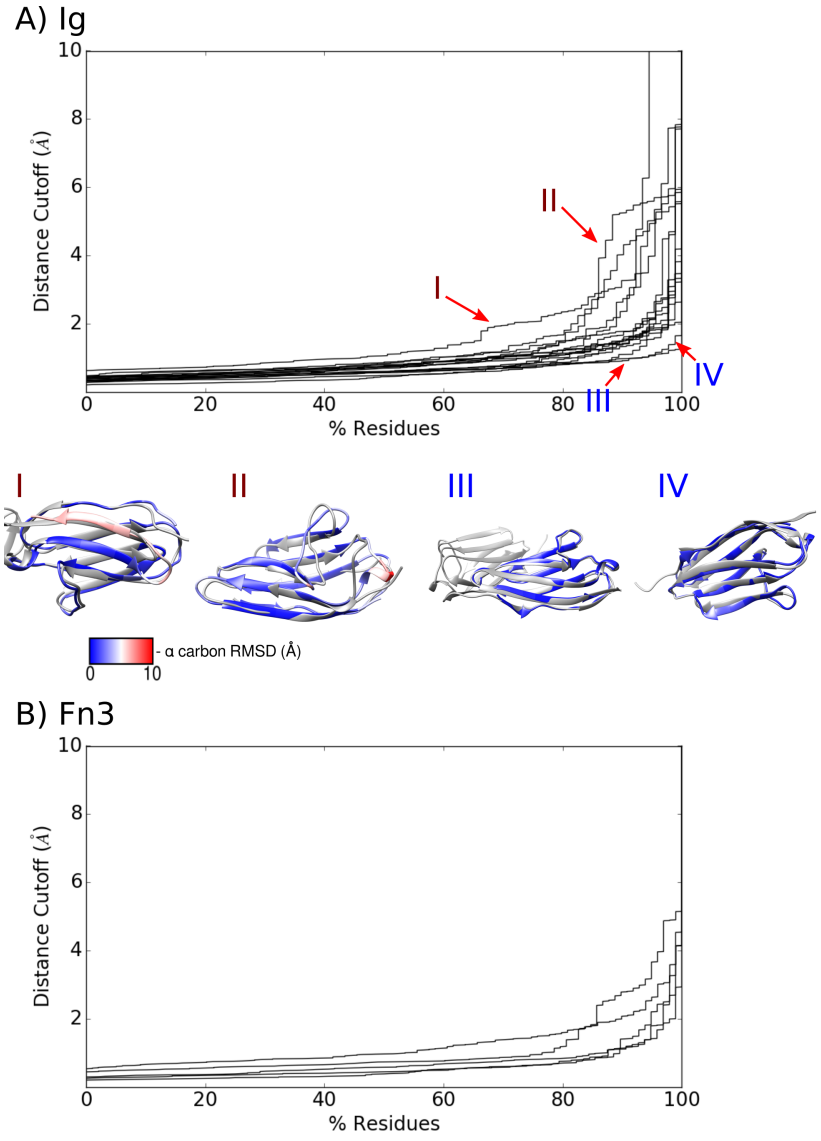

Figure S7: Cumulative RMSD plots for models aligned to representative structures **A** for Ig domains **B** for Fn3 domains. Alignments of Ig models with representative structures coloured by RMSD can be seen for two of the least successful cases (**I** Ig-19 aligned to 5JDD and **II** Ig-164 aligned to 1TNN) and two of the most successful cases (**III** Ig-2 aligned to 2A38 and **IV** Ig-163 aligned to 3QP3). A similar method of model assessment has been used by Sánchez and Sali (1998) and Cozzetto et al. (2009)

```
T-COFFEE, Version_11.00.8cbe486 2014-08-12 22:05:29 - Revision 8cbe486 - Build 477
  Cedric Notredame
  CPU TIME:0 sec.
  SCORE=938
  *
   BAD  AVG  GOOD
  *
  Ig-19     :   91
  2yuzA     :   91
  2dltA     :   91
  1x44A     :   91
  cons      :   93

  Ig-19     ---LHITKTMKNIEVPETKTASFECEVSHFNVPSMWLKNGVEIEM
  2yuzA     SSGLKILTPLTDQTVNLGKEICLKCEISE-NIPGKWTKNGLPVQE
  2dltA     SGQLEVLQDIADLTVKAAEQAVFKCEVSDEKVTGKWYKNGVEVRP
  1x44A     SSGIMVTKQLEDTTAYCGERVELECEVSEDDANVKWFKNGEEIIP

  cons        : :   : :  .   :   ::**:*. .   * ***  :


  Ig-19     --SEKFKIVVQGKLHQLIIMNTSTEDSAEYTFVCGNDQVSATLTV
  2yuzA     --SDRLKVVQKGRIHKLVIANALTEDEGDYVFAPDAYNVTLPAKV
  2dltA     --SKRITISHVGRFHKLVIDDVRPEDEGDYTFVPDGYALSLSAKL
  1x44A     GPKSRYRIRVEGKKHILIIEGATKADAAEYSVMTTGGQSSAKLSV

  cons      ..:   :   *: * *:* ..   * .:* .    :   .:


  Ig-19     T---
  2yuzA     HVIS
  2dltA     NFLE
  1x44A     DLKS

  cons
```

Figure S8: Alignment of query sequence and templates for the domain Ig-19, obtained using T-coffee (Notredame et al., 2000).

References

A. Auton et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393.

L. Cavallo et al. POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res*, 31(13):3364–6, 2003.

C. Chauveau et al. Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Hum Mol Genet*, 23(4): 980–91, 2014a. doi: 10.1093/hmg/ddt494.

C. Chauveau et al. A rising titan: TTN review and mutation update. *Hum Mutat*, 35(9):1046–59, 2014b. doi: 10.1002/humu.22611.

D. Cozzetto et al. Evaluation of template-based models in CASP8 with standard measures. *Proteins*, 77 Suppl 9:18–28, 2009. doi: 10.1002/prot.22561.

R.C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004. doi: 10.1186/1471-2105-5-113.

A. Evilä et al. Atypical phenotypes in titinopathies explained by second titin mutations. *Ann Neurol*, 75(2):230–40, 2014. doi: 10.1002/ana.24102.

A. Evila et al. Targeted Next-Generation Sequencing Reveals Novel TTN Mutations Causing Recessive Distal Titinopathy. *Mol. Neurobiol.*, Oct 2016. doi: 10.1007/s12035-016-0242-3.

R.D. Finn et al. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*, 39(Web Server issue):W29–37, 2011. doi: 10.1093/nar/gkr367.

R.D. Finn et al. Pfam: the protein families database. *Nucleic Acids Res*, 42(Database issue): D222–30, 2014. doi: 10.1093/nar/gkt1223.

A. Ganna et al. Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.*, 19(12): 1563–1565, Dec 2016. doi: 10.1038/nn.4404.

B. Gerull et al. Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat Genet*, 30(2):201–4, 2002. doi: 10.1038/ng815.

A. González-Pérez and N. López-Bigas. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, 88(4):440–9, 2011. doi: 10.1016/j.ajhg.2011.03.004.

P. Hackman et al. Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. *Am J Hum Genet*, 71(3):492–500, 2002. doi: 10.1086/342380.

D.S. Herman et al. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med*, 366(7): 619–28, 2012. doi: 10.1056/NEJMoa1110186.

M. Itoh-Satoh et al. Titin mutations as the molecular basis for dilated cardiomyopathy. *Biochem Biophys Res Commun*, 291(2):385–93, 2002. doi: 10.1006/bbrc.2002.6448.

R. Izumi et al. Exome sequencing identifies a novel TTN mutation in a family with hereditary myopathy with early respiratory failure. *J Hum Genet*, 58(5):259–66, 2013. doi: 10.1038/jhg.2013.9.

M. Lek et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616): 285–91, 2016. doi: 10.1038/nature19057.

X. LIU et al. [Titin gene mutations in Chinese patients with dilated cardiomyopathy]. *Zhonghua Xin Xue Guan Bing Za Zhi*, 36(12):1066–9, 2008.

L.R. Lopes et al. Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *J Med Genet*, 50(4):228–39, 2013. doi: 10.1136/jmedgenet-2012-101270.

B.J. Maron et al. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation*, 92(4):785–9, 1995.

Y. Matsumoto et al. Functional analysis of titin/connectin N2-B mutations found in cardiomyopathy. *J Muscle Res Cell Motil*, 26(6-8):367–74, 2005. doi: 10.1007/s10974-005-9018-5.

S. Mundy et al. Duchenne/becker muscular dystrophy: advances in reproductive testing options. *Fertility and Sterility*, 106(3):e372, 2016. doi: 10.1016/j.fertnstert.2016.07.1058.

E.W. Myers and W. Miller. Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–7, 1988.

C. Notredame et al. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17, 2000. doi: 10.1006/jmbi.2000.4042.

M. Ohlsson et al. Hereditary myopathy with early respiratory failure associated with a mutation in A-band titin. *Brain*, 135(Pt 6):1682–94, 2012. doi: 10.1093/brain/aws103.

O. O'Sullivan et al. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340(2):385–95, 2004. doi: 10.1016/j.jmb.2004.04.058.

J. Palmio et al. Hereditary myopathy with early respiratory failure: occurrence in various populations. *J Neurol Neurosurg Psychiatry*, 85(3):345–53, 2014. doi: 10.1136/jnnp-2013-304965.

A. Pantazis et al. Diagnosis and management of hypertrophic cardiomyopathy. *Echo Res Pract*, 2(1):R45–53, 2015. doi: 10.1530/ERP-15-0007.

G. Pfeffer et al. Titin mutation segregates with hereditary myopathy with early respiratory failure. *Brain*, 135(Pt 6):1695–713, 2012. doi: 10.1093/brain/aws102.

U. Pieper et al. MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*, 37(Database issue):D347–54, 2009. doi: 10.1093/nar/gkn791.

D.E. Pires et al. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–42, 2014a. doi: 10.1093/bioinformatics/btt691.

D.E. Pires et al. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*, 42(Web Server issue):W314–9, 2014b. doi: 10.1093/nar/gku411.

M. Pollazzon et al. The first Italian family with tibial muscular dystrophy caused by a novel titin mutation. *J Neurol*, 257(4):575–9, 2010. doi: 10.1007/s00415-009-5372-3.

A. Porollo et al. Prediction-based fingerprints of protein-protein interactions. *Proteins*, 66(3): 630–45, 2007. doi: 10.1002/prot.21248.

K.D. Pruitt et al. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue):D130–5, 2012. doi: 10.1093/nar/gkr1079.

R. Roncarati et al. Doubly heterozygous LMNA and TTN mutations revealed by exome sequencing in a severe form of dilated cardiomyopathy. *Eur J Hum Genet*, 21(10): 1105–11, 2013. doi: 10.1038/ejhg.2013.16.

R. Sánchez and A. Sali. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A*, 95(23):13597–602, 1998.

S. Schafer et al. Titin-truncating variants affect heart function in disease cohorts and the general population. *Nat. Genet.*, Nov 2016. doi: 10.1038/ng.3719.

C.E. Seidman and J.G. Seidman. Identifying sarcomere gene mutations in hypertrophic cardiomyopathy: a personal history. *Circ Res*, 108(6):743–50, 2011. doi: 10.1161/CIRCRESAHA.110.223834.

M.Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–24, 2006. doi: 10.1110/ps.062416606.

C. M. Strom et al. Cystic fibrosis testing 8 years on: lessons learned from carrier screening and sequencing analysis. *Genet. Med.*, 13(2):166–172, Feb 2011. doi: 10.1097/GIM.0b013e3181fa24c4.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45(D1):D158–D169, Jan 2017. doi: 10.1093/nar/gkw1099.

D.L. Theobald and D.S. Wuttke. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, 22(17):2171–2, 2006. doi: 10.1093/bioinformatics/btl332.

C.M. Topham et al. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng*, 10(1):7–21, 1997.

C. Toro et al. Exome sequencing identifies titin mutations causing hereditary myopathy with early respiratory failure (HMERF) in families of diverse ethnic origins. *BMC Neurol*, 13:29, 2013. doi: 10.1186/1471-2377-13-29.

A. Uruha et al. Necklace cytoplasmic bodies in hereditary myopathy with early respiratory failure. *J Neurol Neurosurg Psychiatry*, 86(5): 483–9, 2015. doi: 10.1136/jnnp-2014-309009.

P.Y. Van den Bergh et al. Tibial muscular dystrophy in a Belgian family. *Ann Neurol*, 54 (2):248–51, 2003. doi: 10.1002/ana.10647.

N. Vasli et al. Next generation sequencing for molecular diagnosis of neuromuscular diseases. *Acta Neuropathol*, 124(2):273–83, 2012. doi: 10.1007/s00401-012-0982-8.

K. Wang et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16): e164, 2010. doi: 10.1093/nar/gkq603.

B. Webb and A. Sali. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci*, 86:2.9.1–2.9.37, 2016. doi: 10.1002/cpps.20.

Y. Zhang. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40, 2008. doi: 10.1186/1471-2105-9-40.

Chapter 3

Missense variants in health and disease impact on distinct functional pathways and proteomics features

3.1 Introduction

In this chapter, we zoom out from our focus on titin variants to perform a large-scale analysis of the properties of missense variants in health and disease. Chapter 2 has shown that titin variants remain difficult to classify, with predictors frequently offering conflicting interpretations. Moreover, the problem of missing heritability, discussed in Chapter 1, remains unsolved, suggesting that a number of variant-disease associations cannot be detected by current statistical methods.

Due to these observations, it becomes increasingly pressing to understand the molecular characteristics of variants in health and disease; including differences in the characteristics of driver and passenger somatic cancer mutations, and in the impact of rare and common population variants. Analyses of the localisation of variants to protein structure, taking into account their proximity to functional sites (e.g. post-translational modifications, or PTMs) (David et al., 2012; Engin et al., 2016; Gao et al., 2015; Gress et al., 2017; Laddach et al., 2018; Lu et al., 2016b), have shown to be effective in uncovering the impact of variants at the molecular level (Pires et al., 2014b). In the field of cancer research, protein structure-based methods have been used to successfully predict cancer driver

genes (Porta-Pardo and Godzik, 2014; Porta-Pardo et al., 2015b), as validated by a recent large-scale study by Bailey et al. (2018). Despite such success, this does not appear to have been applied to other classes of variants (i.e. population and Mendelian disease-associated variants). Only a few studies (Gress et al., 2017; Sivley et al., 2018) have taken advantage of the recently available large-scale data, and compare, using structural bioinformatics methods, disease-associated variants with somatic cancer variants, and variants found in the general population. Furthermore, recent papers propose that previously observed trends, which suggest that somatic cancer single amino acid variants (SAVs) are enriched in protein-protein interaction sites, could be due to biases caused by the tendency of disease-associated variants to localise to those proteins which are most experimentally studied (Gress et al., 2017). Therefore robust statistical methods to treat these comparisons are urgently needed. Additionally, as discussed in Section 1.2.6, large proteomics and transcriptomics datasets have been generated in recent years, but, to the best of our knowledge, studies which incorporate this information in the analysis of the impact of genetic variants are yet to appear.

These factors have motivated us to undertake an integrative analysis which compares disease-associated variants, including somatic cancer variants and germline disease-associated variants, with variants found at different frequencies in the general population. A unique feature of our analysis is in addressing the interplay between atomistic and macroscopic features. Here we define atomistic features to be those features associated with the precise localisation of SAVs to proteins and protein structure, and macroscopic features to be properties of the proteins to which variants localise (e.g. abundance and function). In particular, we have made use of recently published protein half-life data (Franken et al., 2015), along with protein abundance (Wang et al., 2015), thermal stability (Mathieson et al., 2018) and transcriptomics data (GTEx Consortium, 2013), to uncover underexplored biophysical and biochemical principles governing the impact of variants.

It must be noted that Joseph Ng has contributed to the analyses presented in this chapter. Specifically he contributed to the analysis of variant enrichment in oncogenes and tumour suppressor genes presented in Section 3.3.2; performed the DNA-binding domain case study presented in Section 3.3.4; curated the drug-protein mappings used in Section 3.3.4; obtained the gene expression data from the GTEx database, used in Section 3.3.5; and calculated the gene-wise proportion of samples with an RPKM equal to zero (see Section 3.2.1).

3.2 Methods

The analysis of the properties of missense variants in health and disease has necessitated the collection of variant data from available databases, and the mapping of this data to proteins and protein structures/structural homologs. These have been annotated with structural features (e.g. core, surface and interface regions, proximity to post-translational modifications (PTMs)) to allow for the analysis of variant enrichment in protein structural regions. Structural annotations, including residue level mappings of protein-protein interaction sites, have been stored in the ZoomVar database (http://fraternalilab.kcl.ac.uk/ZoomVar). Additionally, a number of other annotations have been collated in order to investigate the properties of variant-enriched proteins/domains, including proteomics, transcriptomics and functional features, as well as protein structural architectures. The following sections describe the collection of this data in detail, along with the methods used to calculate variant enrichment.

3.2.1 Data sources

Variant data

ClinVar (dbSNP BUILD ID 149) variant data (Landrum et al., 2016), COSMIC coding mutations (v80) (Forbes et al., 2015) and gnomAD exome data (Lek et al., 2016), all mapped to the GRCh37 genome build, were obtained in variant call format (VCF). The ClinVar dataset contains variants submitted through clinical channels. Only variants with CLINSIG codes 4 and 5 (probably pathogenic and pathogenic) were selected for further analysis. To ensure the quality of our dataset, we selected only variants with "variant suspect reason code" of 0 (unspecified). Additionally, all variants labelled as being somatic were filtered from this dataset. All variant datasets were mapped to Ensembl protein sequences (Aken et al., 2016) using the Variant Effect Predictor (VEP) (McLaren et al., 2016), and further mapped to canonical UniProt sequences and the respective structures/homologs.

Protein-protein interaction networks

A large non-redundant protein-protein interaction network (UniPPIN) (Chung et al., 2018) was used. This incorporates non-redundant data amalgamated from IntACT (Orchard et al., 2014),

BioGRID (Chatr-Aryamontri et al., 2017), STRING (Szklarczyk et al., 2015), DIP (Xenarios et al., 2002) and HPRD (Peri et al., 2004), as well as recent large-scale experimental studies (Havugimana et al., 2012; Huttlin et al., 2015; Rolland et al., 2014). Although the data from STRING contains computationally predicted interactions, some of which may be false positives, we deem it unlikely that many incorrectly predicted interactions will have structural coverage. As we only analyse those interactions further which can be mapped to structure, we believe our analysis is unlikely to be compromised by false positives in the set of computationally predicted interactions.

Protein sequences and structures

The biounit database of the Protein Data Bank (PDB) was downloaded on 28/04/2017. For mapping purposes, in this study, both the canonical UniProt human protein sequences (Poux et al., 2017) (for mapping to structures and protein-protein interaction networks) and Ensembl protein sequences (Aken et al., 2016) (for mapping variant datasets) were used.

Gene and protein annotations

Gene sets for KEGG pathways were obtained from the MSigDB database (Subramanian et al., 2005). Oncogene and tumour suppressor gene annotations were taken from Supplementary Table S2A from Vogelstein et al. (2013) [1]. Cancer drivers were taken from the Cancer Gene Census (CGC) (COSMIC v84). Genes from both tiers 1 and 2 were included. Conversions between gene symbols, Entrez gene identifiers and UniProt accession numbers were performed using the biomaRt package (Durinck et al., 2005). A list of DNA-binding domains was obtained from the review by Vaquerizas et al. (2009). These domains were mapped from InterPro (Finn et al., 2017) IDs to PFAM IDs using conversion tables in PFAM (v31) [1].

Protein-drug interaction mapping

A mapping of protein-drug interactions was obtained from DrugBank (v5.0.11) (Wishart et al., 2018) (under "Target Drug-UniProt Links") and filtered for human proteins [1]. Drugs were mapped to a PFAM domain-type if at least one domain of that type occurs in a protein a drug is known to interact

[1] Work done by Josef Chi-Fung Ng (Fraternali laboratory)

with. It is, of course, possible that a drug may only interact directly with another domain-type within the protein. However, if only domain-drug interactions with supporting structural information are accepted, the data becomes both sparse and biased towards structurally resolved domains.

Protemics and transcriptomics data

Protein thermal stability and half-life data were obtained from separate large-scale studies by the Savitski lab (Franken et al., 2015; Mathieson et al., 2018). Gene expression quantification (Reads Per Kilobase of transcript per Million mapped reads [RPKM]) counts per sample (v6p) was downloaded from the GTEx portal (GTEx Consortium, 2013) and grouped by tissue, according to the sample metadata provided [1]. For each tissue type, we quantified the gene-wise proportion of samples with an RPKM equal to zero [1]. Only those genes with zero counts in < 10 % of samples were retained for our analysis. Protein abundance data (protein per million [ppm]), integrated for each tissue/sample type were obtained from PaxDb (Wang et al., 2015).

3.2.2 ZoomVar Database

Identification of resolved structures/homologs

Canonical UniProt human protein sequences were assigned resolved structures/homologs from the PDB biounit database (Berman et al., 2003) using BLAST (Altschul et al., 1997). BLAST searches were carried out using both the entire protein sequences and domain sequences, which were defined by scanning UniProt sequences against the PFAM seed library (Finn et al., 2016) using HMMER (Finn et al., 2011). Hits were only accepted with sequence identity > 30 % and E-value < 0.001. T-COFFEE (Notredame et al., 2000) was used to obtain a residue level mapping of queries to structure hits. The quotient solvent accessible surface area [Q(SASA)] of each structure residue was computed using POPS (Cavallo et al., 2003).

Mapping of Ensembl proteins

Ensembl protein sequences were mapped to UniProt protein sequences (Poux et al., 2017), using the UniProt ID mapping. Additionally, if UniProt and Ensembl sequences were not of the same length, the sequences were aligned using T-COFFEE (Notredame et al., 2000) to obtain a per residue

mapping. Stretcher (Myers and Miller, 1988) was used to align those sequences which were too long to align using T-COFFEE.

Identification of interaction complexes

For each interaction in our protein-protein interaction network, resolved binary interaction complexes and homologues were identified using the BLAST search results. As an example, if protein A and B are annotated as interacting in UniPPIN, and their structure homologues A' and B' are located in a resolved structural complex (and at least one residue from each protein is involved in a shared interface), residues from A and B are mapped onto A' and B' to infer their interaction interface.

The partner-specific regression formula from HomPPI (Xue et al., 2011) was used to assign a score and zone to each interaction interface inferred in this way:

$$IC\ Score = \beta_0 + \beta_1 logEVal + \beta_2 PositiveS + \beta_3 Frac_{AA'} + \beta_4 Frac_{BB'} \tag{3.1}$$

The formula uses the following terms:

$$logEVal = \frac{log(EVal_{AA'}) + log(EVal_{BB'})}{2} \tag{3.2}$$

$$PositiveS = \frac{PositiveS_{AA'} + PositiveS_{BB'}}{2} \tag{3.3}$$

$$Frac_{AA'} = Frac_A \cdot Frac_{A'} \tag{3.4}$$

$$Frac_{BB'} = Frac_B \cdot Frac_{B'} \tag{3.5}$$

$$Frac_A = \frac{LAL_{AA'}}{length_A}, Frac'_A = \frac{LAL_{AA'}}{length_{A'}}, Frac_B = \frac{LAL_{BB'}}{length_B}, Frac'_B = \frac{LAL_{BB'}}{length_{B'}} \quad (3.6)$$

Here $EVal_{AA'}$ and $EVal_{BB'}$ represent the BLAST E-values between protein A and A', and between B and B' respectively; $PositiveS_{AA'}$ and $PositiveS_{BB'}$ are the analogous BLAST positive scores. Similarly $LAL_{AA'}$ and $LAL_{BB'}$ are the lengths of the BLAST alignments.

The parameters β_0, β_1, β_2, β_3 and β_4 have values -0.505, 0.001, 0.009, 0.341 and 0.205 respectively. Zone boundaries defined by HomPPI are given in Table 3.1.

Zone	Property	Threshold
Safe Zone	$PositiveS$	$\geq$ 70 %
	$Frac_{AA'}$	$\geq$ 80 %
	$Frac_{BB'}$	$\geq$ 80 %
	$logEVal$	$\leq$ -50
Twilight Zone 1	$PositiveS$	$\geq$ 60 %
	$Frac_{AA'}$	$\geq$ 60 %
	$Frac_{BB'}$	$\geq$ 60 %
	$logEVal$	$\leq$ 1
Twilight Zone 2	$PositiveS$	$\geq$ 55 %
	$Frac_{AA'}$	$\geq$ 40 %
	$Frac_{BB'}$	$\geq$ 40 %
	$logEVal$	$\leq$ 1
Dark Zone	$PositiveS$	$\geq$ 0 %
	$Frac_{AA'}$	$\geq$ 0 %
	$Frac_{BB'}$	$\geq$ 0 %
	$logEVal$	$\leq$ 1

Table 3.1 Zone boundaries for the inference of protein-protein interaction interfaces defined by HomPPI (Xue et al., 2011).

Residues involved in interfaces were assigned using POPSCOMP (Kleinjung and Fraternali, 2005). Only those residues with a change in SASA $> 15\text{Å}^2$ were annotated as interface residues. Many residues contribute only a very small surface area to the interface, and are thus likely to play a less important role in binding affinity and specificity. Setting the threshold to (greater than) 15Å^2 retrieves $>$ 90 % of the interface surface area (91.1 % calculated from the PDB biounit database) and removes residues with negligible individual contributions to this.

Determination of per-residue binding partners

A protein may interact with multiple other proteins. For each of these interactions, a maximum of 10 corresponding best hits (ordered by HomPPI defined score (Xue et al., 2011)), located in the best populated zone, were considered. If a residue was located at the interaction interface, in at least half of these structures, it was annotated as interacting with that specific protein, otherwise it was annotated as non-interacting. This follows the HomPPI procedure to identify interface residues (Xue et al., 2011).

Mapping of variant data

Variants in each dataset were annotated according to protein region localisation using the ZoomVar database. For certain analyses the COSMIC data was divided into "driver" and "non-driver" subsets, taking drivers as variants which map to all proteins from both tier 1 and tier 2 of the Cancer Gene Census (CGC) (COSMIC v84). The non-driver subset contains all other variants.

Definition of regions

We defined several types of protein and domain regions as described below.

Interface regions were considered to be composed of residues which bind to at least one protein interaction partner. Core residues were defined as those with a Q(SASA) < 0.15. Surface residues were defined as those with a Q(SASA) ≥ 0.15 which do not take part in protein-protein interaction interfaces. These thresholds are identical to those used by Stehr et al. (2011).

Disordered protein regions were predicted using DISOPRED3 (Jones and Cozzetto, 2015). Intra-domain ordered regions were defined as those regions predicted to be ordered which lie within PFAM defined domains. Intra-domain disordered regions were defined as regions predicted to be disordered which lie within PFAM domains. Inter-domain disordered regions were defined as those regions not located within PFAM defined domains which are predicted to be disordered. Any residues with structural coverage were filtered from the inter-domain disordered regions as it was thought that these could potentially belong to domains which have not been defined by PFAM. Although DISOPRED3 lacks sensitivity (0.147) it has high specificity (0.958) (Jones and Cozzetto, 2015). Therefore we have high confidence in the annotation of those regions which were predicted

to be disordered. Furthermore, due to the large size of our dataset, we believe that this procedure allowed for trends in the variant enrichment of disordered and ordered protein regions to be captured, despite noise which was inevitably introduced due to prediction error.

Ubiquitination and phosphorylation sites were obtained from PhosphoSitePlus (Hornbeck et al., 2004). Each site was mapped to the structural template with the highest identity. Regions close to phosphorylation and ubiquitination sites were defined as those within 8 Å in 3D space. A cut-off of 8 Å has been frequently been used to define regions close to functional sites, for example by Stehr et al. (2011).

Creation of ZoomVar database

All data, including per-residue mappings, were stored in the ZoomVar MySQL database (MySQL, 2008). A web interface and REST architecture was implemented, using the Django framework (Django Software Foundation, 2017) to allow users to query this. It is available at http://fraternalilab.kcl.ac.uk/ZoomVar.

3.2.3 Calculation of SAV enrichment

The binomial cumulative distributive function (see Equation 3.7) was used to assess the SAV enrichments of individual proteins, domains or domain-types, and the 2-tailed binomial test was used to assess the significance of enrichment/depletion. In this formula k is the number of observed SAVs which localise to a region, n is the total number of SAVs which localise to all regions of interest (all_regions) and p is ratio of the size of the region (number of amino acids) to the size of all_regions:

$$P(N(SAVs_{region}) \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{3.7}$$

Hereafter, we refer to the binomial CDF as the variant enrichment score (VES).

The calculations were performed for the regions defined in Table 3.2. For each analysis at the whole protein or whole domain level, all UniProt proteins/domains (except for immunoglobulin and T cell receptors), which contained SAVs in any of the datasets analysed, were considered to be the

background proteome (all_regions). Proteins belonging to immunoglobulin and T cell receptor gene family products were filtered from all analyses (HGNC definition (Yates et al., 2017)), to avoid the inclusion of variants which could have arisen from the process of affinity maturation.

level/region	all_regions
protein	all proteins
domain	all domains
PFAM domain-type	all PFAM domain-types
protein region	union of all regions of a particular protein
domain region	union of all regions of a particular domain
PFAM domain-type region	union of all regions of a particular PFAM domain-type

Table 3.2 The anatomy of the protein levels considered in our analysis. N.B. at the protein region, domain region and domain-type region levels, if the region of interest is the ”core”, the union of all regions will be the ”core”, ”surface” and ”interface”. The same logic applies to other regions of interest (see Fig. 3.1).

For all calculations of enrichment and simulations involving protein or domain regions (e.g. core, surface and interface), only those proteins/domains which posses the region of interest and to which SAVs localise (although not necessarily to the region of interest), were considered. For example, to calculate the enrichment of SAVs at the core of a protein, the protein must have a core region of size > 0. Moreover, a region of a protein cannot be enriched in SAVs, in comparison to the rest of the protein, if no SAVs from a particular dataset localise to that protein.

The overall SAV enrichment of protein regions, for each data set, was also calculated using a density-based metric (see Equation 3.8).

$$P(SAV_{region}) = \frac{(N(SAVs)_{region}/size_{region})}{(N(SAVs)_{all\_regions}/size_{all\_regions})} \tag{3.8}$$

Here 95 % confidence intervals were estimated via bootstrapping (10,000 iterations). The 2-tailed significance of enrichment/depletion was estimated by simulation. 10,000 simulations were carried out for each dataset, in which the number of variants which localise to a given protein was kept constant, but their location within the protein randomised. The regional density of variants was calculated for each simulation and compared to the actual value in order to derive a p-value. Simulations were performed in this way, keeping the number of SAVs which localise to each protein

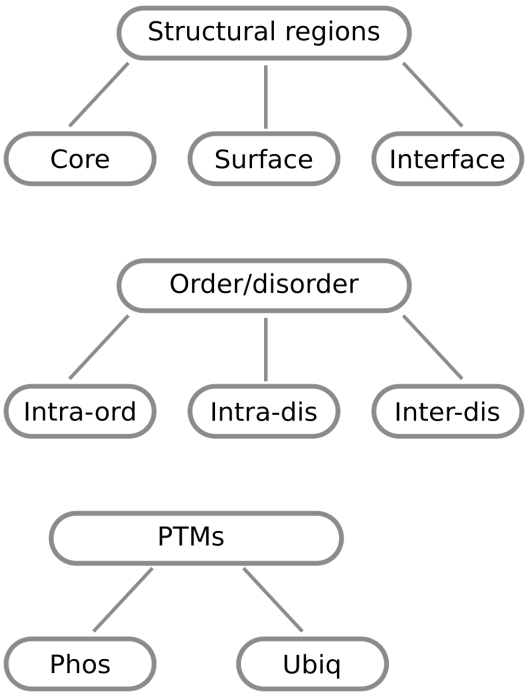

Fig. 3.1 The regions used for the analysis of variant enrichment at the protein region, domain region and domain-type region levels (as defined in Table 3.2)

fixed, in order to overcome bias which stems from the assumption that variants are uniformly distributed throughout the proteome.

3.2.4 Further protein structural analyses

Calculation of protein topological network features

Graph representations for protein structures, with mapped SAVs, were constructed. Here protein Cα atoms are represented by nodes in the graph, and nodes are connected if Cαs are < 10 Å apart in 3D space. Similar graph representations are routinely used to create elastic network models of proteins. One class of such models is the Gaussian network model (GNM); the default cut-off distance used by GNMs is commonly 10 Å (Bakan et al., 2011). Therefore we use this cut-off in the creation networks here. Topological network features were calculated using the python package NetworkX (Hagberg et al., 2008). Specifically, we calculate the degree, degree centrality, betweenness centrality and closeness centrality of residues (nodes) to which variants localise. The degree of a node is defined as the number of nodes it is connected to, and the fraction of connected nodes constitutes its degree centrality. Betweenness centrality BC is defined as the number of pairwise shortest paths which pass through a node:

$$BC(u) = \sum s, t \in V \frac{\sigma(s, t|u)}{\sigma(s, t)} \tag{3.9}$$

Here V is the set of nodes, $\sigma(s, t)$ is the total number of shortest paths, and $\sigma(s, t|u)$ is the number of shortest paths which pass through the node u.

Closeness centrality CC is the reciprocal distance of the sum of the shortest paths from all other $n - 1$ nodes to node v:

$$CC(u) = \frac{n - 1}{\sum_{v=1}^{n-1} d(v, u)} \tag{3.10}$$

Where $d(v, u)$ is the distance of the shortest path between nodes v and u.

Calculation of protein core density

The density of residue packing in protein cores may impact on their ability to accommodate missense variants. To investigate this, a non-redundant set of representative structures for the UniProt canonical proteome was collated. Here, structures from the ZoomVar database were mapped in order of identity. Additional structures were only added to represent a protein if at least 50 % of the residues they covered were not mapped to a structure with higher identity. Protein core density was only calculated for those proteins with a single mapped structure. This was to avoid problems associated with determining the core density for multi-domain proteins. The number of Cα contacts ($<$ 8 Å) for each core residue [Q(SASA) $<$ 0.15] was counted. The mean number of contacts for all core residues within a protein was used as a proxy for protein density, the assumption being that protein cores with greater density will have a higher number of Cα contacts. Only protein cores which comprise of $>$ 4 residues were analysed, as it was considered density could not be accurately calculated for very small cores. It was also considered that if this threshold was set too high, the data would become too sparse to observe trends. However, the results presented in Section 3.3.5 remained consistent for all cut-offs between $>$ 2 and $>$ 8 residues, with only small differences seen outside these cut-offs (see Appendix A1).

3.2.5 Enrichment analysis of gene sets and domain architecture sets

Gene set enrichment

Enrichment analyses were performed using Gene Set Enrichment Analysis, using the implementation provided by the R FGSEA package (Sergushichev, 2016). Given an enrichment statistic for each query gene, the GSEA algorithm outputs a score per gene set, which quantifies the enrichment of query genes in the sets examined. This is then normalised by the size of the gene set, to give a normalised enrichment score (NES).

We utilise the centred VES, as the enrichment statistic which is input into the GSEA algorithm. Here, the centred VES is simply obtained by subtracting 0.5, therefore proteins with the expected number of SAVs have a centred VES of 0. At the whole protein level only sets with n $\geq$ 25 were considered. At the protein subregion level, variant enrichment data exists for a smaller number of

proteins, due to incomplete structural coverage of the proteome. In order to perform a complete comparison between pathway enrichment at different levels, all pathways analysed at the whole protein level were also analysed at the protein subregion level.

Definition of pathway clusters

The pathway normalised enrichment scores (NESs), calculated at the whole protein level for each dataset, were used to perform K-means clustering of KEGG pathways (Kanehisa et al., 2017). The R package NbClust (Charrad et al., 2014) was used to determine the optimum number of clusters.

CATH architecture enrichment analysis

PFAM domains (Finn et al., 2016) were assigned CATH domains (Sillitoe et al., 2015) using the per-residue mapping of structures to domains, from both data sets, available from the SIFTs resource (Velankar et al., 2013). A minimum of 50 residues, which mapped to both a particular CATH domain and a particular PFAM domain were required to assign a PFAM domain to a CATH domain. This threshold was used in order to prevent spurious assignments. If a PFAM domain appeared to map to more than one CATH domain, the majority vote, from the residue level mapping, was used. Using these assignments PFAM domains were mapped to CATH architectures, guided by the CATH hierarchy, to create "domain sets" for each architecture.

Architecture enrichment analysis was performed as for gene set enrichment analysis, however here the domain-type, and domain-type region centred VESs (for each PFAM domain-type) were used as the enrichment statistics. Additionally, as we were interested in the enrichment of individual architectures only "domain" sets of size $n \geq 25$ were considered at all levels.

3.2.6 Analysis of expression, abundance, density, and stability data

Spearman correlations of protein-wise and region SAV enrichments with expression levels (RPKM), abundance (ppm), half-life (hours), thermal stability (Tm), and density (mean contacts of core α carbons) were calculated. Additionally, gene set enrichment analysis was performed as in Section 3.2.5, but using the metrics in Table 3.3 as enrichment statistics.

thermal stability	Tm - mean(Tm)
abundance	log(ppm + 1) - mean(log(ppm + 1))
expression	log(median(RPKM) + 1) - mean(log(median(RPKM) + 1))
half-life	log(hours) - mean(log(hours))

Table 3.3 Proteomics and transcriptomics-based metrics used as enrichment statistics for GSEA analysis.

Here it can be seen that the mean value for each quantity of interest was subtracted to obtain values centred around 0, allowing both pathway enrichment and depletion to be assessed.

3.2.7 Statistics and data visualisation

The majority of data analyses were performed in the R statistical programming environment. All corrections for multiple testing have been done using the Benjamini-Hochberg method in R (p.adjust function). Bootstrapping was performed using the boot package (function boot) (Canty and Ripley, 2017). Spearman correlations were performed using the SpearmanRho function of the DescTools package (Signorell, 2017). Heatmaps were produced with either the heatmap.2 function in the gplots package (Warnes et al., 2016) or the ComplexHeatmap package (Gu et al., 2016), in which clustering, wherever shown, was performed with hierarchical clustering (hclust function) using default parameters unless otherwise stated. Circos plots were generated with the Circos package (Krzywinski et al., 2009). Additionally, binomial CDFs were calculated and two-tailed binomial tests performed using the NumPy package in Python (Oliphant, 2006).

3.3 Results

We present a multidimensional analysis of single amino acid variants (SAVs) observed in the general population (gnomAD database) (Lek et al., 2016), in comparison to somatic cancer-associated SAVs from the COSMIC database (Forbes et al., 2015) and disease-associated SAVs from the ClinVar database (Landrum et al., 2016). Throughout this analysis we further divide the gnomAD data into its constituent common (MAF ≥ 0.01) and rare (MAF < 0.01) variants, to investigate whether there are differences between these two subsets.

We ask whether the enrichment of variants is associated with specific structural features and functional pathways, and whether results differ for population and disease-associated variants. In particular we investigate the interplay between variant enrichment and proteomics features; for example, we explore whether disease-associated variants are found more frequently in the cores of less thermally stable proteins, as these might be more easily sufficiently destabilised to lead to complete/partial unfolding. Such exploration of the interplay of these atomistic features with macroscopic features is novel in the field. Finally, we use these features to understand whether rare population variants demonstrate characteristics which are more similar to common population variants or disease-associated variants.

Our analysis explores the enrichment of SAVs at different levels, constituting what we define as a protein-centric anatomy of variants in health and disease, as illustrated in Fig. 3.2. We employ a similar approach to that used in the prediction of cancer driver genes (Porta-Pardo et al., 2015a): the SAV enrichment of individual proteins/regions has been modelled using a binomial distribution (Methods Equation 3.7), whereas global trends in the distribution of SAVs have been investigated by calculating SAV density (Methods Equation 3.8). The binomial cumulative distribution function quantifies the enrichment of variants (Fig. 3.2g) and is referred to as the Variant Enrichment Score (VES). This is assessed statistically using a two-tailed test (see Section 3.2.3). Additionally, the significance of the enrichment/depletion of SAVs, in terms of their density, is assessed by comparison to simulated SAV distributions, in which the number of SAVs is kept identical to that observed in the data, but their positions within the protein are randomised. This goes beyond similar studies (e.g. (David et al., 2012; Engin et al., 2016; Gao et al., 2015)) and addresses possible biases in our current knowledge of protein structures and interactions.

A summary of the numbers of SAVs investigated in each dataset is given in Table 3.4, and a more detailed breakdown is given in Appendix A2.

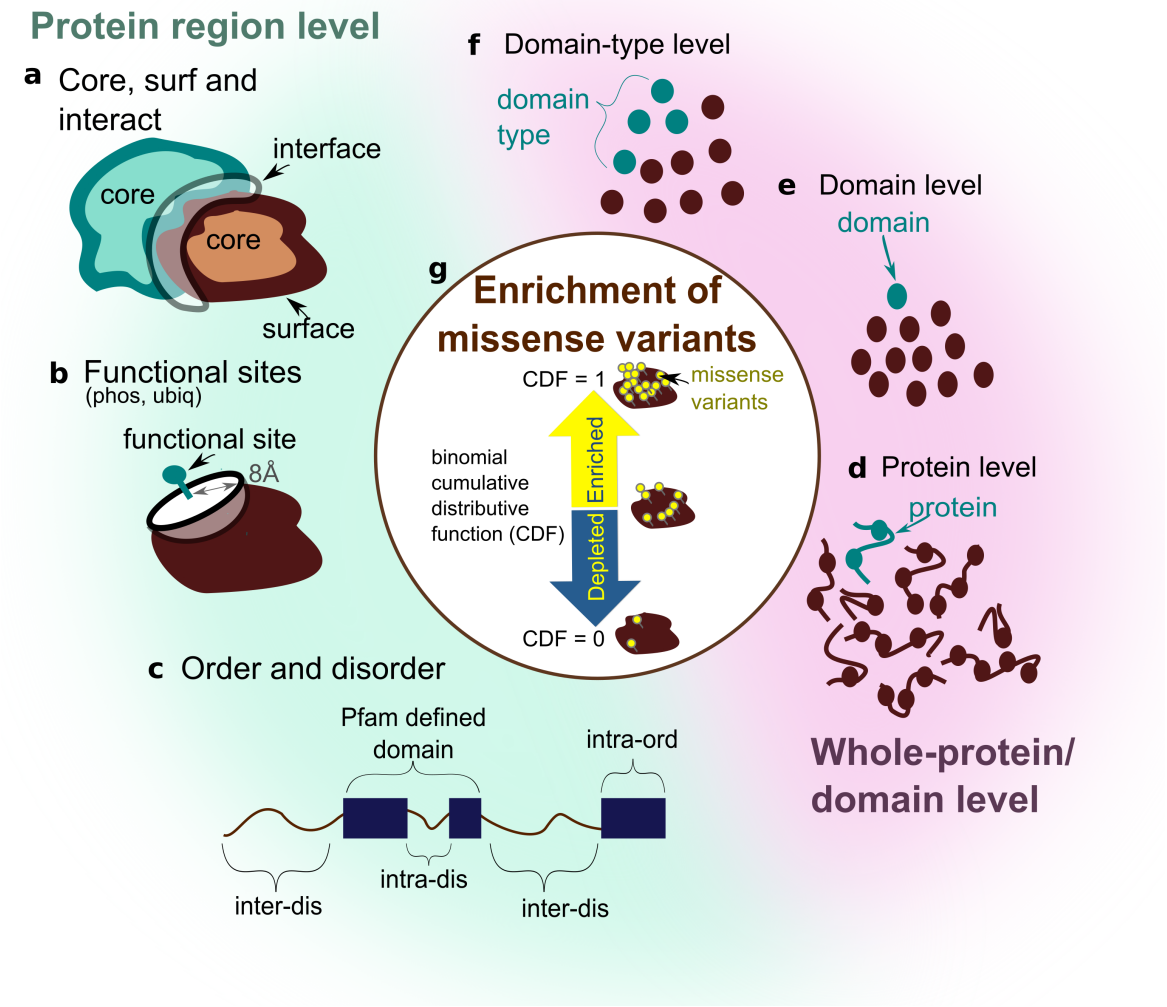

Fig. 3.2 Enrichment statistics are calculated at different levels. At the protein region level, the number of SAVs in a region is compared to the number of SAVs in the rest of the protein. Regions are defined as: a) core, surface and interface regions of a protein; b) regions close to functional sites, and; c) regions predicted to be ordered or disordered which lie either within or outside of PFAM defined domains. d,e) At the protein/domain level the number of SAVs in a protein or domain is compared to the number of SAVs in the whole dataset which localise to defined proteins/domains. f) At the domain-type level the number of SAVs in a particular PFAM defined domain-type is compared to the number of SAVs which localise to all domains. g) The calculation of enrichment at the different levels is statistically assessed using the binomial distribution. The binomial cumulative distributive function constitutes a Variant Enrichment Statistic (VES) with value range 0 to 1, which quantifies enrichment.

Region	Common	Rare	COSMIC	ClinVar
protein	54571	3806698	1731030	21272
surf	12151	966409	491179	8558
interact	403	38108	22205	768
core	2789	296291	152356	5194
intra-ord	20650	1575286	755683	14620
intra-dis	2984	197482	96914	1211
inter-dis	17352	1045997	439437	1128
phos	1661	158192	82364	2362
ubiq	440	52250	25778	607

Table 3.4 Numbers of SAVs which localise to different protein regions in the studied datasets. Data shown for common and rare population variants from the gnomAD database, somatic cancer variants from the COSMIC database and disease-associated variants from the ClinVar database.

3.3.1 Disease-associated and population variants impact on different functional pathways

We first investigated whether variants from each dataset impact on proteins which are involved in distinct functional pathways. To do this we performed KEGG (Subramanian et al., 2005) functional pathway analysis, by ranking proteins using their whole-protein VESs (see Fig. 3.2d) calculated for each dataset, and using the Gene Set Enrichment Analysis (GSEA) algorithm (Subramanian et al., 2005) (see Section 3.2.5).

The pathway enrichment data, for each mutation dataset, were subjected to clustering and Principal Component Analysis (PCA) (see Section 3.2.5). In Fig. 3.3a it can be seen that variant enrichment segregates pathways into three clusters. Strikingly each pathway cluster appears to have distinct characteristics. The cluster visualised in orange is primarily composed of terms associated with cancer, growth and proliferation, whereas that coloured pink contains pathways associated with splicing, transcription, translation and metabolic terms. Pathways associated with sensory perception and the immune response are found in the final "green" cluster. A handful of metabolic pathways also localise to this cluster, however, these appear to be more associated with environmental response and adaptation than those pathways found in the "pink" cluster; for example, pathways associated with the metabolism of drugs and xenobiotics are found here. For brevity, the "orange", "pink" and "green" clusters will be termed the "proliferative", "nucleotide processing" and "response" clusters respectively, for the remainder of this text. A list of pathways assigned to each cluster is given in Appendix A3.

This visualisation (Fig. 3.3a) also reveals that both the common and rare subsets of the gnomAD database associate tendentially with the "response" cluster, whereas COSMIC data localises between the clusters associated with response and proliferation. ClinVar data associates with the "nucleotide processing" cluster, between both the "response" and "proliferation" clusters. Strikingly, the population variant datasets (gnomAD rare and common) are clearly separated from the disease-associated variant datasets by the first principal component (PC1), whereas COSMIC variants are separated from ClinVar variants along the third principal component (PC3) (see Fig. 3.4).

These trends of functional distinction are further visualised in the Circos plot (Fig. 3.3b). Here it can be clearly seen that the gnomAD data only shows significant enrichment for pathways belonging to the "response" cluster, whereas the COSMIC data shows enrichment for pathways belonging to this cluster and those belonging to the "proliferative" cluster. The ClinVar dataset displays enrichment for pathways belonging to all three clusters; uniquely showing enrichment for pathways within the "nucleotide processing" cluster.

We went on to extend this analysis to the protein region level (Fig. 3.5). Here we find that proteins enriched in gnomAD variants at the surface (Fig. 3.3c) are significantly enriched in pathways belonging to the "proliferative" cluster. Moreover, this enrichment is shared between common and rare variants (albeit not significant for common variants in individual pathways after FDR correction). Proteins with surfaces enriched in disease-associated variants (from COSMIC and ClinVar) are, contrastingly, not enriched in "proliferative" cluster pathways. However, no such pattern emerges for the protein core and interface (Fig. 3.3d), possibly suggesting that population variants avoid disrupting the function of proliferation-related proteins by preferentially localising to the surface. Interestingly, the "nucleotide processing" cluster does not show such a marked enrichment of variants which localise to the surface in the gnomAD database, perhaps indicating that these pathways are more robust to disruption than those in the proliferative cluster. These data show that there is clearly an interplay between variant localisation at the macroscopic level (functional pathways) and the atomistic level (structural regions).

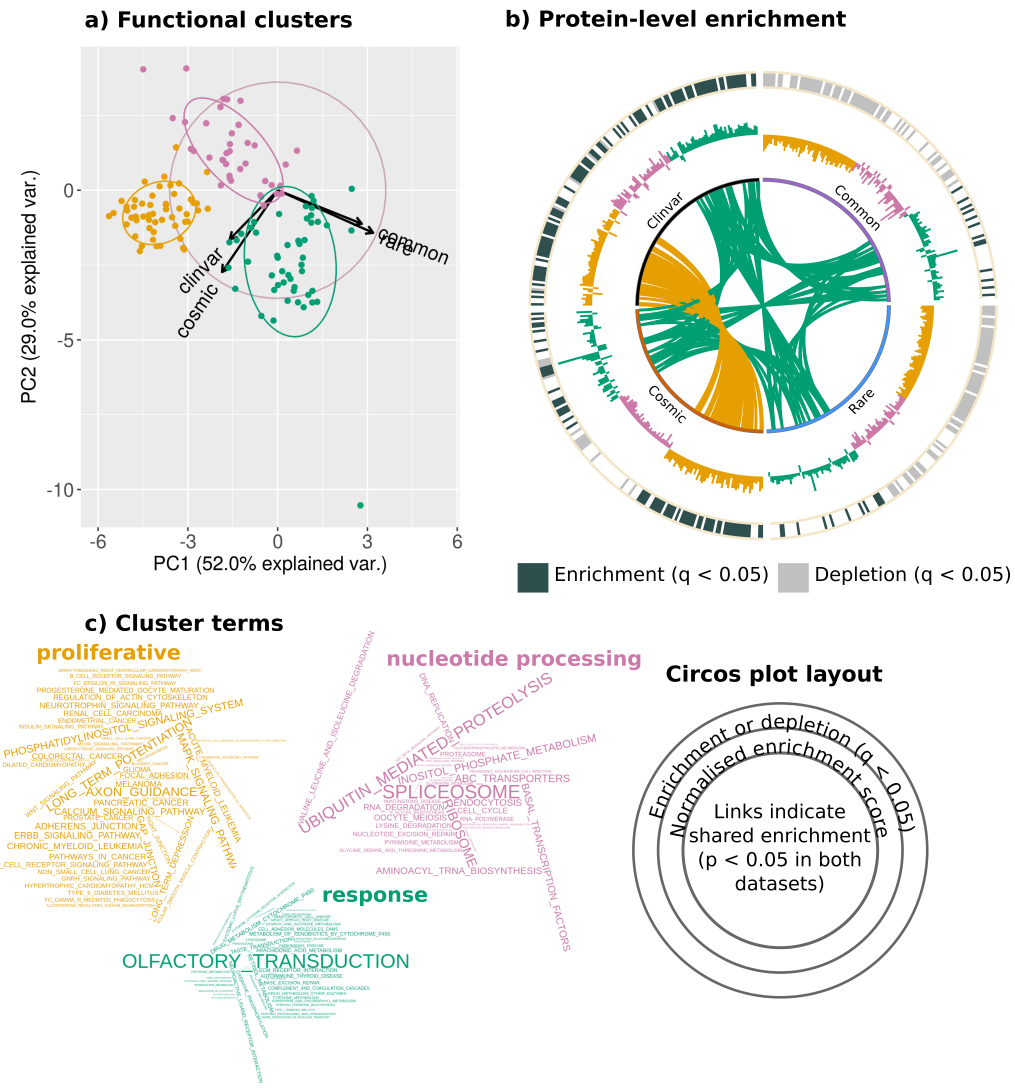

Fig. 3.3 Functional analysis of proteins according to variant enrichment. Gene set enrichment for KEGG functional pathways was performed by ranking proteins using their whole protein VESs, which quantify the variant enrichment of a protein in comparison to the whole proteome. The analysis was performed separately for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). a) At the whole protein level, KEGG pathways form 3 identifiable clusters (K-means), as visualised projected on the first two principal components of the PCA. Each cluster has been assigned a colour for visualisation purposes; these colours correspond to those used in b) and c). COSMIC, ClinVar and gnomAD (rare/common) data can be clearly separated by pathway enrichment, as evidenced by the visualisation of factor loadings (arrows). b) Enrichment for each dataset, at the whole protein level (as in a), visualised on a Circos plot, with results for each dataset occupying a quarter of the plot. Pathways are coloured and ordered by cluster membership defined in a) and c). The Normalised Enrichment Score for each pathway is plotted as a bar graph (the further from the centre, the more positive) in the middle layer of the plot. In the outermost layer of the plot, significant enrichment (dark grey) or depletion (light grey) of a pathway (q-value < 0.05) is depicted. In the centre of the plot, links indicate enrichment (p-value < 0.05) shared between datasets. c) Pathway terms visualised by cluster, sized by their cluster uniqueness score. This is defined as the average of the Euclidian distances (calculated in 4D) to the two other cluster centres. Each cluster has been titled ("proliferative", "nucleotide processing" or "response") to reflect its pathway composition.

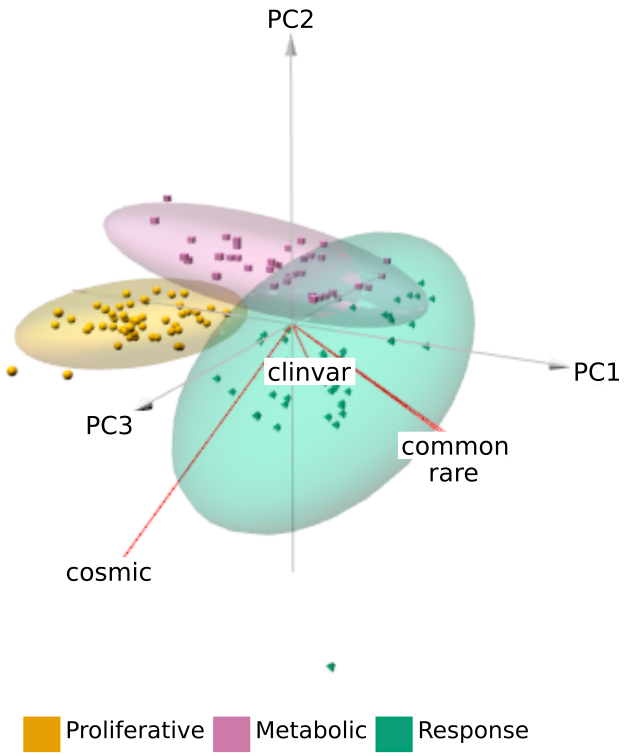

Fig. 3.4 Functional analysis of proteins according to variant enrichment. Gene set enrichment, for KEGG functional pathways, was performed by ranking proteins using their whole protein VESs, which quantify the variant enrichment of a protein compared to the whole proteome. The analysis was performed separately for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). A 3D PCA representation shows that at the whole protein level KEGG pathways form 3 identifiable clusters (K-means), as projected onto the first three principal components. Each cluster has been assigned a colour for visualisation purposes, and has been titled ("proliferative", "nucleotide processing" or "response") to reflect its pathway composition. It can be seen that the gnomAD rare/common data are clearly separated from disease-associated data (cosmic/clinvar) by the first principal component, whereas cosmic and clinvar data are separated by the third principal component.

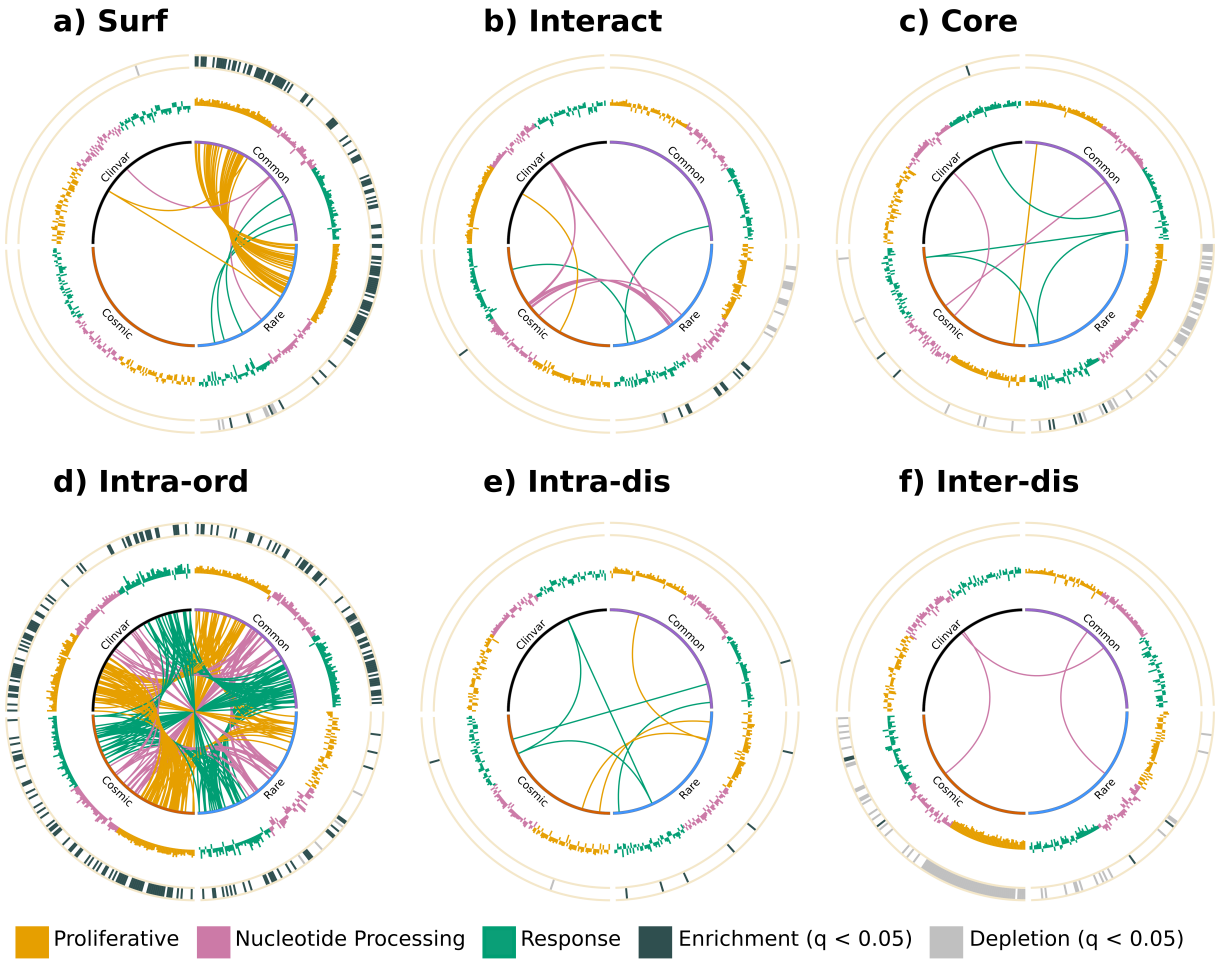

Fig. 3.5 Functional analysis of proteins according to variant enrichment at different region levels (a-f), visualised on a Circos plot. Gene set enrichment for KEGG functional pathways was performed by ranking proteins using their protein region VESs, which quantify the variant enrichment of a protein region in comparison to the whole protein. The analysis was performed separately for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). Results for each dataset occupy a quarter of the plot. The normalised enrichment score for each pathway is plotted as a bar graph (the further from the centre, the more positive) in the middle layer of the plot. In the outermost layer of the plot, significant enrichment (dark grey) or depletion (light grey) of a pathway (q-value < 0.05) is depicted. In the centre of the plot, links indicate enrichment (p-value < 0.05) shared between datasets. Pathways are coloured and ordered by cluster membership, as defined in 3.3. Each cluster has been titled ("proliferative", "nucleotide processing" or "response") to reflect its pathway composition.

3.3.2 Population and disease-associated variants localise to different protein regions

We then zoomed in to view trends in the enrichment of variants at the atomistic level. Specifically, we catalogued the enrichment of variants in core, surface and interface regions; intra-domain ordered regions (intra-ord), intra-domain disordered regions (intra-dis), and inter-domain disordered regions (inter-dis); and regions close to ($\leq$ 8Å) of phosphorylation sites and ubiquitination sites.

In agreement with previous research, we find disease-associated (ClinVar) variants to be enriched in both protein cores and interfaces, but depleted on protein surfaces (see Fig. 3.6a) (David et al., 2012; Engin et al., 2016; Gao et al., 2015; Gress et al., 2017). This reflects the disruption, caused by such mutations, of structurally and functionally important protein regions. GnomAD variants (both common and rare) and somatic non-driver variants display the opposite trend, most likely as variants which localise to protein surfaces are less likely to impact on protein structure and function than either core or interface mutations. Somatic driver variants follow trends closer to ClinVar variants, with slight, but significant, depletion on the surface, but enrichment in the core. Protein interfaces are enriched in disease-associated variants but depleted of gnomAD rare variants. GnomAD common variants appear neither significantly enriched nor depleted, however this may result from the relative sparsity of the data; fewer variants are shared between many individuals (this is clearly evidenced by the numbers in Table 3.4). Interestingly, COSMIC non-driver variants appear depleted in interacting interfaces. However, it becomes clear that they are actually significantly enriched when compared to simulated null distributions (Fig. 3.6a). It is likely that a small number of proteins which harbour a large number of variants at interface regions dominate this variant set, as evidenced by the overlaps of 95 % confidence intervals of the observed and simulated distributions; these may be putative cancer driver genes (see Appendix A4), as a number of known driver genes are enriched in variants in protein interface regions (Bailey et al., 2018; Engin et al., 2016; Porta-Pardo et al., 2015a). Indeed, the enrichment of variants in such regions has been used by Porta-Pardo et al. (2015a) to identify cancer driver genes.

A more detailed per-protein analysis can bring finer granularity into the comparison of variant enrichment. Therefore we look at a curated list of oncogenes and tumour suppressor genes (TSGs)

(see Section 3.2.1) (Vogelstein et al., 2013). As stated in Section 3.1, these analyses were performed in collaboration with Joseph Chi-Fung Ng (Fraternali laboratory).

Several studies have suggested that proteins encoded by oncogenes (which are activated upon mutation) and tumour-suppressor genes (TSGs, which are inactivated upon mutation) tend to be enriched in mutations in different protein regions (Engin et al., 2016; Gress et al., 2017; Stehr et al., 2011). We found that clustering based on VESs broadly classifies these proteins into two groups, one comprising of proteins enriched in mutations mainly at protein-protein interaction interfaces and protein surfaces, and another group of proteins generally enriched in mutations in the core (some of these proteins also show enrichment in mutations in interacting interfaces but, nonetheless, a clear depletion at the surface is evident) (Fig. 3.6b). Interestingly, we observe a statistically significant (Fisher-exact test p-value = 0.004199) segregation of these two groups in terms of cancer driver status: the first group of proteins are mainly (17 out of 24) products of oncogenes, and the other mainly those of TSGs (17 out of 25). These results are consistent with the hypotheses that activating mutations in oncogenes are likely to affect particular functions by impacting on specific interactions, whilst inactivating mutations in TSGs abrogate protein function (Engin et al., 2016; Stehr et al., 2011). Taking the oncogenes and TSGs as two separate groups, the GSEA result confirms a similar trend; moreover, it can also be seen that the disease-associated datasets (ClinVar and COSMIC) show opposite patterns of enrichment in comparison to the gnomAD data (see Fig. 3.7) (Engin et al., 2016; Gress et al., 2017; Stehr et al., 2011). These results confirm that our approach reproduces previous results and highlights clear robust trends.

On analysis of variant enrichment in ordered and disordered regions, we again observe clear segregation between disease and population variants (see Fig. 3.6a). ClinVar and COSMIC variants are depleted in inter-domain disordered regions and enriched in intra-domain ordered regions. In contrast, gnomAD variants (both rare and common) appear enriched in inter-domain disordered regions and depleted in intra-domain ordered regions. GnomAD common and rare variants show similar trends to one another, which are distinct to those of disease-associated variants. These results suggest that variants are more likely to be pathogenic if they fall within ordered domain regions.

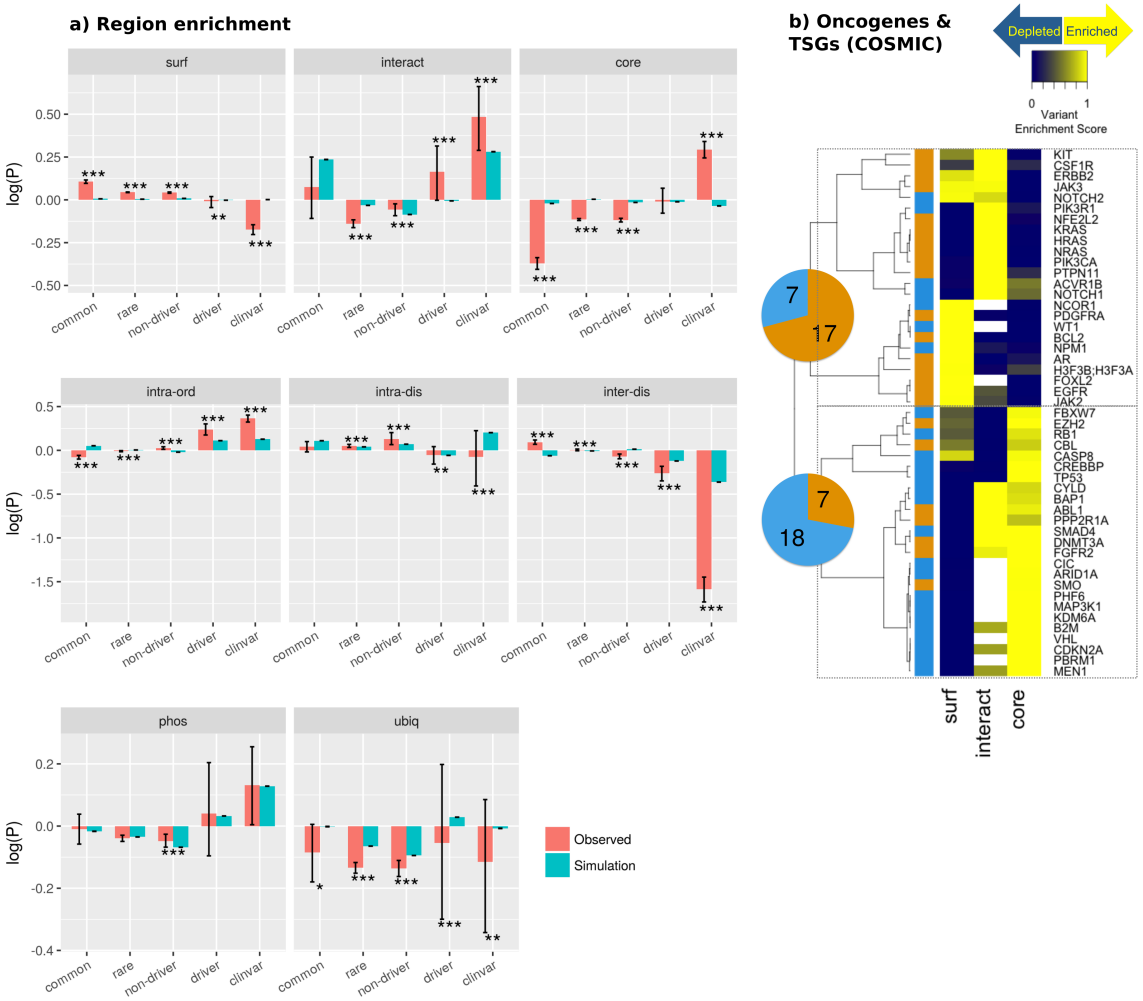

Fig. 3.6 The localisation of missense variants to protein regions. a) The density of missense variants in different protein regions. Shown for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). Observed densities (pink), and densities derived from simulated null distributions (turquoise) are shown. Error bars depict 95 % confidence intervals; for observed densities, these were obtained by bootstrapping. Significance was calculated by comparison of observed values to simulated null SAV distributions (significance level indicated by: * q-value $<$ 0.05, ** q-value $<$ 0.001, *** q-value $<$ 0.0001). b) Enrichment of cosmic missense variants in protein core, surface and interface regions, across a list of annotated oncogene and tumour suppressor gene (TSG) products. The genes were grouped into two clusters using hierarchical clustering (see dendrogram by rows), with the pie charts enumerating the number of oncogenes (orange) and TSGs (blue) in each cluster.

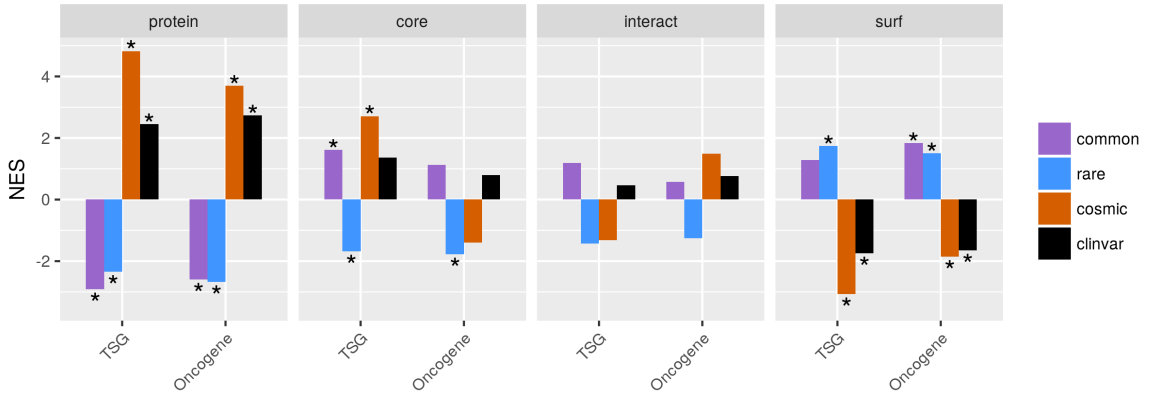

Fig. 3.7 The enrichment of tumour suppressor and oncogene gene sets in common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). Enrichment calculated using the GSEA algorithm using the Variant Enrichment Score (VES) as the enrichment statistic. The VES is calculated to describe the variant enrichment of a protein in comparison to the entire proteome, and the variant enrichment of a protein region (core, surf or interact) in comparison to the rest of a protein. Tumour suppressor gene and oncogene gene sets were defined as described in Section 3.2.1.

The density of variants close to PTMs is also shown in Fig. 3.6a. Here, ClinVar variants appear enriched when considering the density of SAVs close to phosphorylation sites, but not significantly so in comparison to simulations. The large bootstrap confidence interval suggests this may be due to the sparsity of the data available. A similar observation is seen for COSMIC driver variants; however, COSMIC non-driver variants, which appear depleted according to variant density, are significantly enriched close to phosphorylation sites in comparison to simulated null distributions. This indicates that, in agreement with a number of other studies (Olow et al., 2016; Reimand et al., 2013), the disruption of phosphorylation sites may play a particularly important role in cancer. In contrast to phosphorylation sites, all data sets appear depleted of variants close to ubiquitination sites.

These analyses conclude that the enrichment of missense variants at various structural features consistently segregate population variants from disease-associated ones. For the majority of structural regions defined here, the greatest, most consistent distinction is always seen between common and ClinVar variants, given that the data is not too sparse.

3.3.3 Population and disease-associated variants have different topological structural network properties

Here we ask whether variants from different datasets have distinct topological properties according to their structural localisation. We define these topological properties by representing protein structures at networks, in which nodes consist of Cαs, and those Cαs within 10 Å of one another are connected by edges. Due to this representation, we are able to calculate topological network properties of the residues to which variants localise. Such properties give insight into the connectivity and neighbourhood of the affected residues.

The results of the analysis are depicted in Fig. 3.8. Given that we find protein cores to be enriched in ClinVar variants, it is perhaps unsurprising that we find these variants to localise to residues with a significantly higher degree (are more highly connected), than do variants from the other datasets (pairwise Mann-Whitney test, see Appendix A5 for values). We also find that disease-associated variants show significantly higher values for several other centrality measures. Interestingly, the most significant difference between datasets (Kruskal-Wallis test, see Appendix A5), after degree, is betweenness centrality. This metric can be seen to highlight bottlenecks within the network, as it is a measure of the fraction of shortest paths which pass through a node. This suggests that disease-associated variants may impact on residues which play an important role in communication through the structural network. Conversely, disease-associated (ClinVar) variants localise to residues with only slightly higher median degree centrality (the fraction of connected nodes) than population variants, and slightly lower median degree centrality than COSMIC driver variants. This suggests that disease-associated variants localise to residues with a higher degree as they occur in proteins in which all residues are more highly connected. Although we see significant differences between the datasets, it is clear there is a large overlap in their distributions of topological network features (see Fig. 3.8).

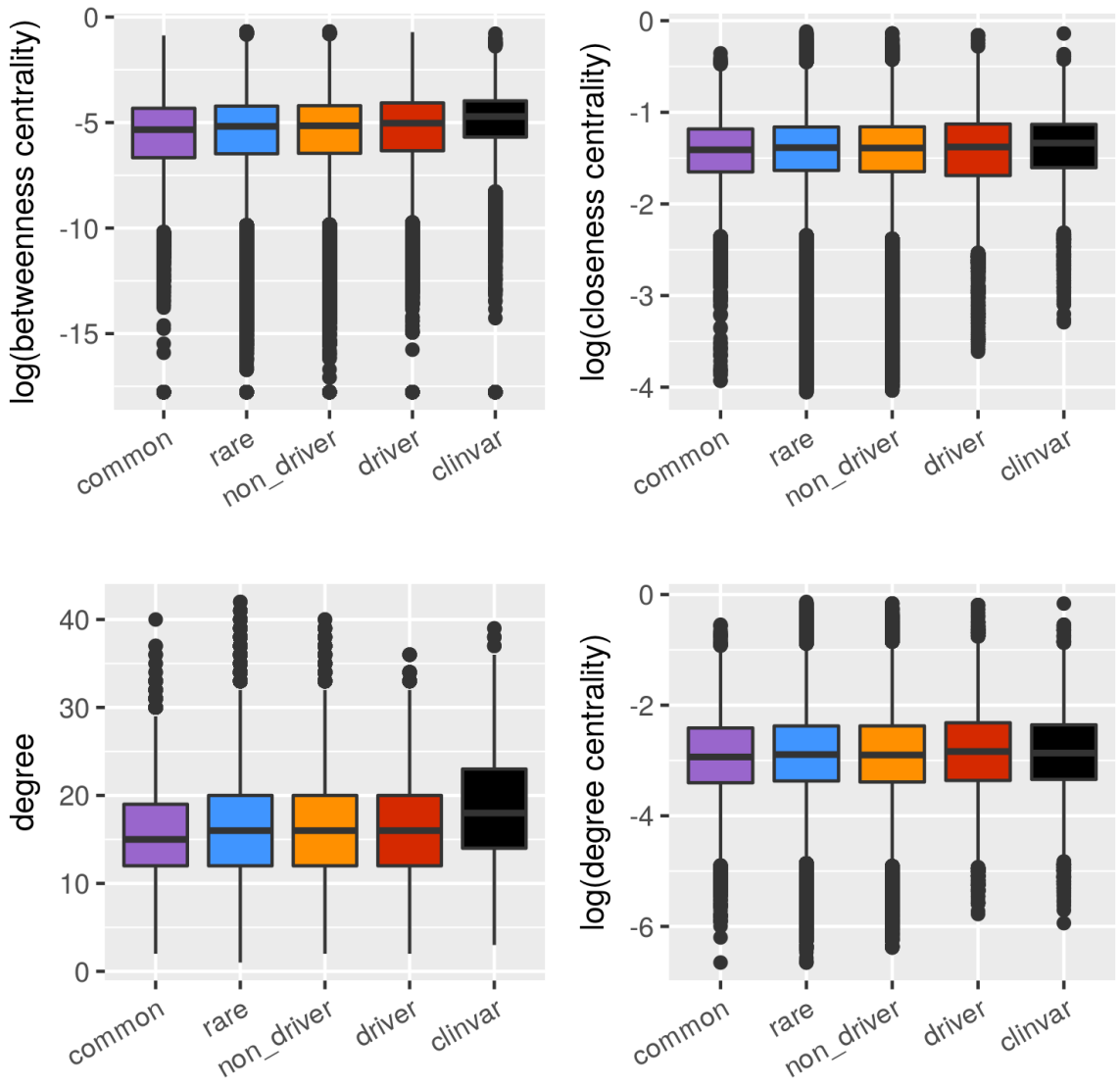

Fig. 3.8 Cα structural network topological features of variants. Proteins are transformed into networks based on positions of Cα carbons. Network topological features for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar) datasets are compared. Somatic cancer variants have been further divided into variants which localise to proteins which are classified as drivers or non-drivers. See Appendix A5 for associated statistics.

3.3.4 Towards a domain-centric landscape of variant enrichment

We then proceeded from the protein level to examine variant enrichment at the domain level. First, we studied whether disease-associated SAVs would localise preferentially to domains with specific architectures. We made use of the CATH protein domain classification system (Sillitoe et al., 2015)

and focussed on the architectural level, which groups domains with similar secondary structural orientations, thereby capturing tertiary structural features. We mapped PFAM domain definitions to those used by CATH, and created domain sets (analogous to gene sets) for each CATH architecture. Enrichment was calculated at both the domain-type level (i.e. localisation of SAVs to a domain-type, for example fibronectin type-III (Fn3), in comparison to localisation of SAVs to all other domain-types, see Fig. 3.2), and at the domain-type region level (i.e. localisation of SAVs to core residues within a domain-type, for example all Fn3 core residues, in comparison to the localisation of SAVs to all other residues within a domain-type).

As depicted in Fig. 3.9, the results show that, at the whole domain level, the data sets show similar trends in variant localisation. A number of architectures, such as the Alpha Horseshoe architecture, show depletion of variants in all data sets, in contrast to other architectures, for example the Beta Sandwich and Irregular architectures, which show enrichment in all data sets. Few architectures, such as the Alpha-Beta Barrel, which is enriched in the gnomAD rare data but depleted for the other data sets, show markedly different patterns of enrichment at the domain architecture level. At the protein region level, the picture diversifies with the ClinVar data generally showing enrichment in architecture cores, although not significantly, in contrast to the other data sets which are more frequently depleted of variants in this region. Interestingly, this trend is particularly marked for the Alpha-Beta Barrel architecture. Thus, although gnomAD rare variants are enriched in this architecture, it is clear that very few of these localise to this architecture's core.

We moved on to compare variant localisation across protein domain-types, using PFAM domain definitions. As depicted in Fig. 3.2, we calculated the variant enrichment at the amalgamated whole-domain and region (core/surface/interaction sites) levels. A case study, performed by Joseph Chi-Fung Ng (Fraternali lab), highlights striking patterns of variant enrichment. Here we focus on DNA-binding proteins: a wealth of analyses have established the fundamental role of structural properties in the interaction of these proteins with DNA (Luscombe and Thornton, 2002; Rohs et al., 2010; Schneider et al., 2014). We considered a list of DNA-binding domains (DBDs) curated in the literature (Vaquerizas et al., 2009), and compared their whole-domain and region VESs. Fig. 3.10 shows that DBDs vary considerably in their enrichment of disease-associated and population variants. Visualisation of the whole-domain level enrichment statistics shows that some DBDs

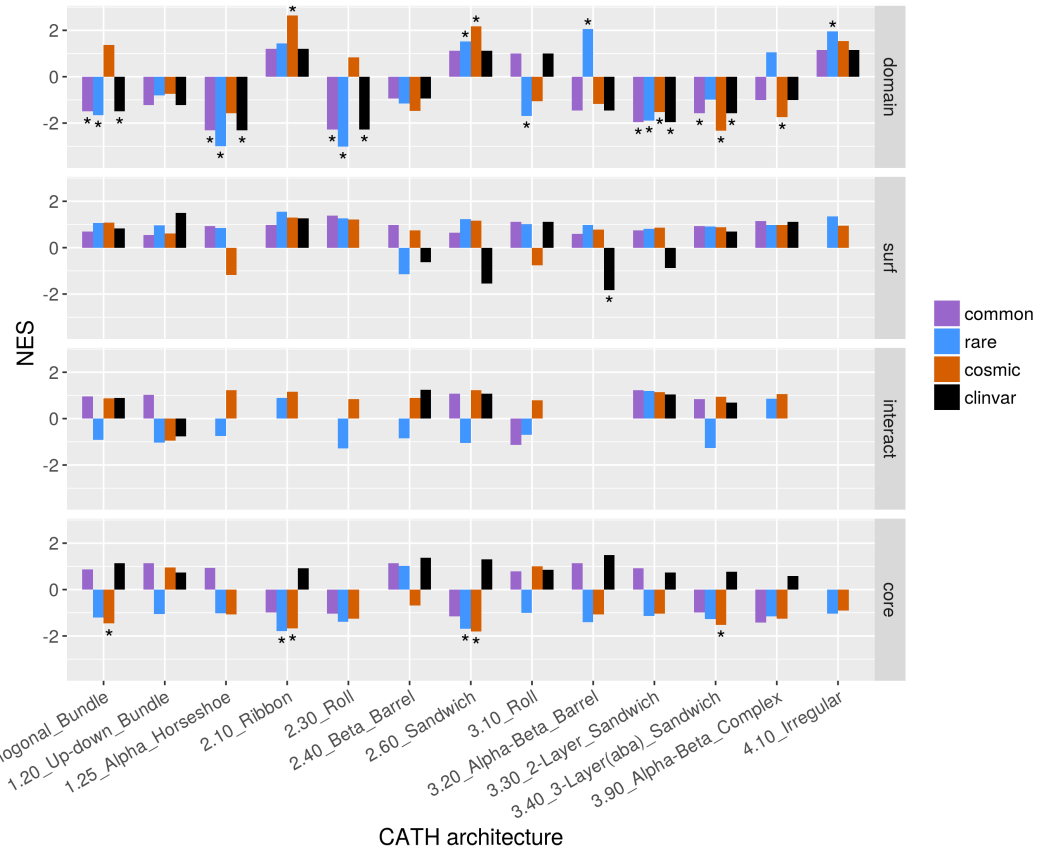

Fig. 3.9 The enrichment CATH architectures in PFAM domains according to SAV enrichment, quantified as the variant enrichment score (VES). Here CATH architectures constitute domain sets, analogous to gene sets, and the VES constitutes the enrichment statistic. The normalised enrichment score (NES) for each architecture is shown on the y-axis of the plots. Results are depicted at the domain level (the enrichment of a PFAM domain in comparison to all other PFAM domains) and the domain region level (the enrichment of the core, surface or interface of a PFAM domain in comparison to the rest of a PFAM domain). * indicates q-value < 0.05.

(e.g. Forkhead, Homeobox) are only enriched in COSMIC and ClinVar variants, while zf-H2C2_2 domains, which are numerous in zinc finger (ZNF) proteins, appear to be enriched only in rare variants. Other DBDs, e.g. Myb_DNA_binding, appear depleted in all types of variants. Different regional enrichment patterns are also observed: ClinVar mutations are typically enriched in the core but devoid at the surface of these domains, whereas in the COSMIC data, as well as the two nominally healthy datasets derived from the gnomAD database, variants which localise to the surface are more common (Fig. 3.10).

Characteristic patterns of variant enrichment also hold when we examine across all PFAM domains, including, but not limited to, DBDs. Fig. 3.11 depicts the union of the top 20 most variant-enriched domains for each data set. Here it can be seen that a small number of domains appear enriched in variants in primarily only the COSMIC and ClinVar data sets. These include known drug targets such as kinase and ion channel domains. A handful of domains, which are only enriched in COSMIC variants, include the Cadherin_tail and Laminin_G_2 domain, both of which are important in cancer (Garg et al., 2014; Jeanes et al., 2008). A larger number of domains are variant enriched in both the COSMIC and gnomAD dataset (rare and common variants). Some domains (e.g. Serpin, UDPGT, Collagen and EGF_CA) contain variants from all datasets or all datasets with the exception of COSMIC. In such domains, it is likely that the precise structural localisation of a variant determines whether it plays a pathogenic role. Intriguingly a few domain-types, such as NPIP and NUT appear only enriched in common variants. This could suggest that these domains take part in functions for which it is desirable to maintain diversity within a population; however, little is known about either domain-type (PFAM, 2018a,b). Thereby the bias in study towards those domains associated with disease, rather than those enriched in population variants, is further highlighted.

It also becomes apparent that the global trends in variant localisation to the core, surface and interface regions, observed in Section 3.3.2 are recapitulated here. Again the majority domains are enriched in gnomAD (rare and common) variants at the surface but ClinVar variants at the core. Although COSMIC variants show a trend broadly similar to gnomAD variants, it is clear that a larger proportion of domain-types are enriched at the core or interface. These include domain-types with known cancer driver associations, such as the P53 and VHL domains (Semenza, 2006).

We wished to understand how the targeting of domains by drugs mapped to the landscape of variant enrichment we observed. To investigate this we used the protein-drug mapping provided in the DrugBank database, as detailed in Section 3.2.1. The curation of this list was performed by Joseph Chi-Fung Ng. As already extensively pointed out (Santos et al., 2017), the targeting of domain-types by existing drugs is highly biased towards a small number of domain-types, such as GPCRs and kinase domains. Indeed, we observe a large number of drugs targeting proteins containing 7tm (GPCR) domains. These domains are enriched in variants from the gnomAD and COSMIC database but are devoid of disease-associated ClinVar variants. Interestingly it has recently been shown that

genetic variants in such domains (GPCRs), identified in the general population, may be associated with differential drug response between individuals (Hauser et al., 2018). Therefore we show that our domain-centric landscape of variant localisation highlights, for each domain-type, implications useful for both understanding variant impact and motivating therapeutic design (see Section 3.4).

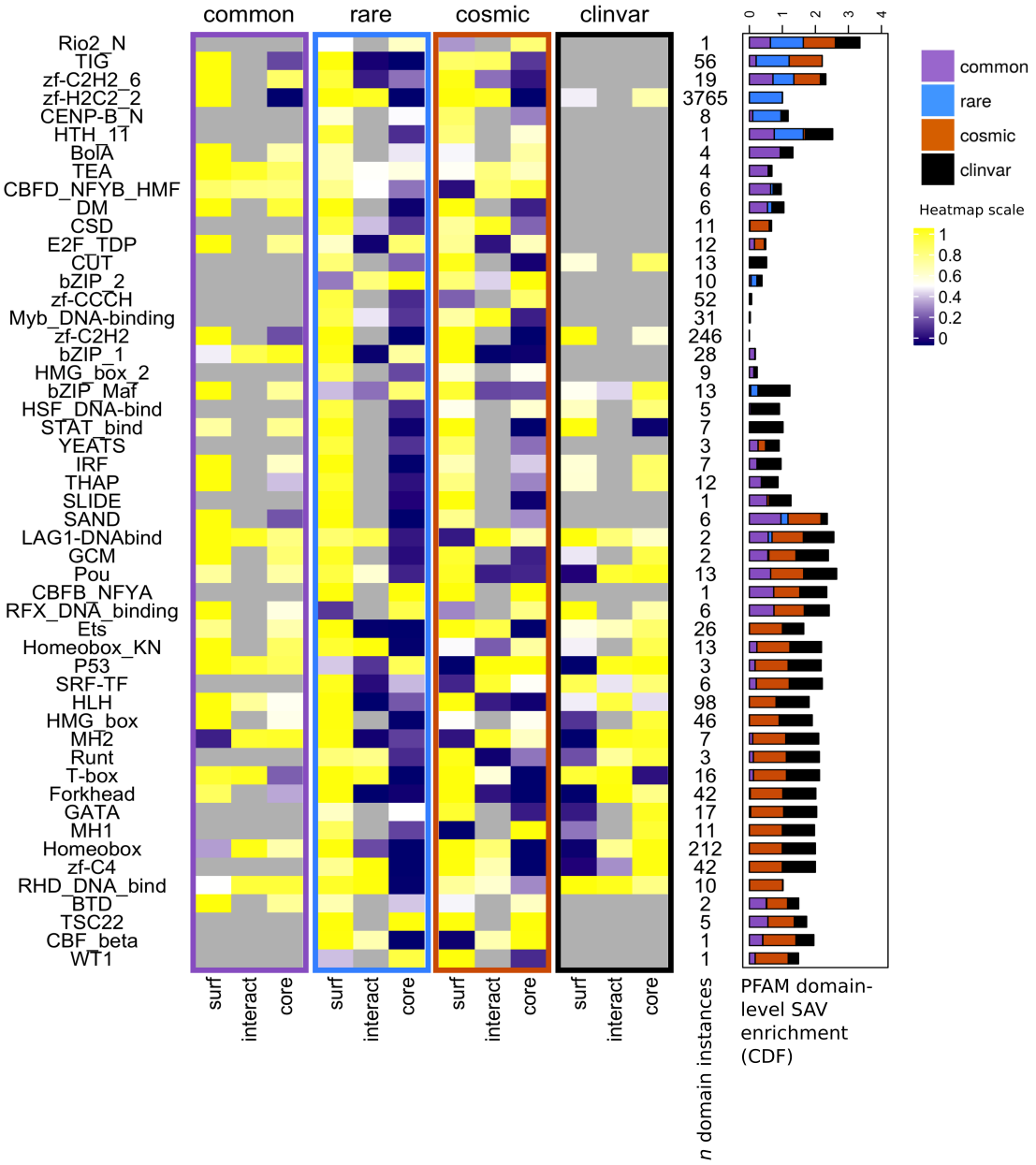

Fig. 3.10 Landscape of variant enrichment in DNA-binding domains (DBDs). Here all DBDs, as curated in Vaquerizas et al. (2009), with structural coverage are considered, and their domain-region level enrichment is depicted. Each row corresponds to a PFAM domain-type. Region (surface, interface and core) enrichments (which compare the enrichment of a region from a PFAM domain-type in comparison to other regions from that domain-type) are shown in the heatmaps for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). Grey cells correspond to the absence of interaction site mappings (for interface data), or the absence of mutations in our dataset. Each row is annotated with the number of domain instances of each type present within our data. A stacked bar graph shows the enrichment at the whole domain level (the variant enrichment of a PFAM domain-type in comparison to all PFAM domains) for each data set.

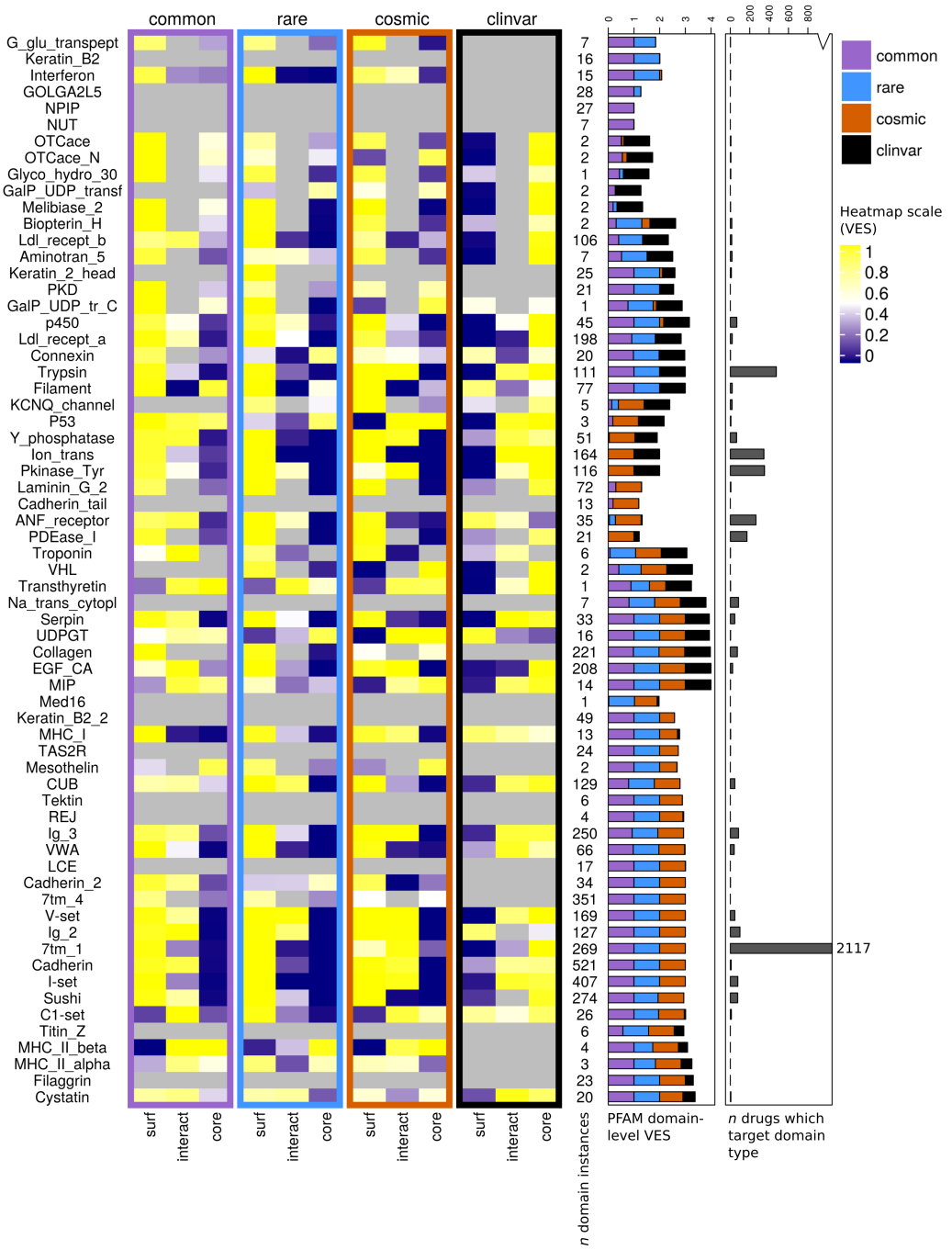

Fig. 3.11 A domain-centric landscape of variant enrichment. Here the union of the top 20 most enriched domains for each data set is depicted: each row corresponds to a PFAM domain-type. The plots show, from the left to the right, heatmaps of regional (surface, interface and core) enrichments for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). These regional enrichments compare the variant enrichment of a region from a PFAM domain-type in comparison to other regions from that domain-type. Grey cells correspond to the absence of interaction site mappings (for interface data), or the absence of mutations in our dataset. Domain instances and enrichment at the whole domain level (the variant enrichment of a PFAM domain-type in comparison to all other domain-types) for each data set are shown as stacked bars to the left of the plot. The number of drugs known to target proteins containing each domain-type is depicted in the rightmost bar graph. Note the cut numeric axis; the number of drugs which target the only outlier, the 7tm_1 domain, is noted on the plot.

3.3.5 Proteomics and transcriptomics features associate with variant localisation

Proteins, of course, do not function in isolation but in the crowded environment of the cell. In our analysis so far we have viewed proteins through their three dimensional and functional properties; however, we have to consider that proteins may be present in the cell in different quantities, display different turnover rates and possess different melting temperatures. All of these factors can crucially affect the stability and the fitness of a protein to perform its function. Here we have made use of large-scale proteomics data, including protein abundance data from PaxDb (Wang et al., 2015) and data describing both protein half-lives and thermal stability from the Savitski lab (Franken et al., 2015; Mathieson et al., 2018), together with transcriptomics data from the GTEx database (GTEx Consortium, 2013), to explore relationships between these features and variant localisation. Please note that the numbers of proteins and SAVs which underlie each comparison are described in Appendix A6.

Our results show that the protein-wise enrichment of disease-associated variants displays positive correlations with protein abundance, expression, half-life and thermal stability, whereas population variants exhibit the opposite trend (see Figs 3.12-3.14). It is important to recall here that the protein-wise enrichment of variants is calculated in comparison to the entire proteome (all UniProt proteins which contain SAVs in any of the datasets; see Fig. 3.2d).

However, if we zoom into the enrichment of variants in the core of protein structures, in comparison to all regions of proteins with resolved structure, rare population variants demonstrate a positive correlation with abundance and thermal stability, whereas disease-associated variants negatively correlate with this (see Fig. 3.12). These results prove robust across multiple tissue types. Analogous correlations for variant enrichments at protein surfaces display roughly opposite trends to those observed at the protein cores. Due to the relative sparsity of variants which map to protein interfaces, we believe it is difficult to draw robust conclusions from any trends observed for correlations of proteomics data with variant enrichment at protein-protein interaction sites.

Our results, at the "core" region level, for gnomAD rare and ClinVar variants suggest that disease-associated variants might preferentially localise to the core of unstable proteins, as these

might be more easily destabilised to a degree at which function is deleteriously impacted. This possibility is further explored in the discussion. Similarly to the ClinVar data, the gnomAD common data also show negative correlations for variants occurring at the protein core; this could potentially give weight to the argument presented by Mahlich et al. (2017) that common variants could affect molecular function more than rare variants. However, we believe this is more likely to be due to the fact that very few common variants localise to protein cores, as shown by Fig. 3.6, resulting in sparse statistics (i.e. the correlation is calculated over Variant Enrichment Scores which are already very low).

One might expect that mutations would be less easily accommodated in cores of densely packed proteins, which would have higher thermal stability. To assess this we calculate the mean number of Cα contacts within 8 Å of core residues, as a proxy for protein density. We find a significant correlation between this metric and protein thermal stability (vehicle_1: ρ = 0.1680595, q-value = 1.464451e-12; vehicle_2: ρ = 0.1854869, q-value = 1.528586e-13). If we correlate this metric of core density with the core Variant Enrichment Score, we find a significant negative correlation for the gnomAD common dataset. No other datasets show significant correlations with core density, however a clear trend emerges in which correlations become progressively more positive in the order of gnomAD common, gnomAD rare, somatic non-driver, somatic driver and ClinVar (see Fig. 3.15). This suggests variants may be more deleterious if they localise to a packed core. Again the complexity of the interplay between features is highlighted, as the higher stability of proteins with more packed cores suggests that destabilisation, to a degree which is physiologically relevant, may be more difficult to achieve. Although core packing and thermal stability are correlated, the correlation value (ρ) is low. Therefore, this feature is clearly not the only determinant of protein stability.

The results we see at the whole protein level, where the disease-associated ClinVar data clearly show a more positive correlation with Tm, are, at a first sight, more difficult to explain. However, work by the Picotti lab (Leuenberger et al., 2017) has demonstrated that more stable proteins are generally more abundant. In agreement with this, we find significant correlations between the protein abundance and thermal stability data (see Appendix A7). Moreover, we do see significant positive correlations of protein-wise variant enrichment with protein abundance, in our analysis

(see Fig. 3.12b). This suggests that the preferential localisation of ClinVar variants to more stable proteins could be attributed to the higher abundance of such proteins.

Interestingly, it can be seen that the trends observed at both the protein level and core region level, are less pronounced for cell line data and break down for extracellular fluids (saliva and urine). Moreover, the trend is most evident for tissues containing long-lived cell-types, such as the brain, ovary and testis. Transcriptomics data (see Fig. 3.13) again reinforces this picture, albeit with less contrast between data sets (particularly at the protein core).

Finally, we wanted to understand whether correlations with these proteomic and transcriptomic features could be associated with the specific functional roles of the involved proteins. This was achieved by investigating the association of these proteomic and transcriptomic features with biological pathways, using the GSEA algorithm. For the majority of proteomic and transcriptomic features, no clear associations with the functional clusters identified in Fig. 3.3 can be detected (see Appendix A8). An exception to this is protein thermal stability: pathways which belong to the "proliferative" cluster are clearly enriched in proteins of lower stability than the other two clusters (see Fig. 3.10c). This suggests that proliferation-related proteins may be vulnerable to disruption by mutations which localise to their already unstable cores. Moreover, this agrees with the idea proposed in Section 3.3.1, that "proliferative" cluster proteins may be less robust to disruption.

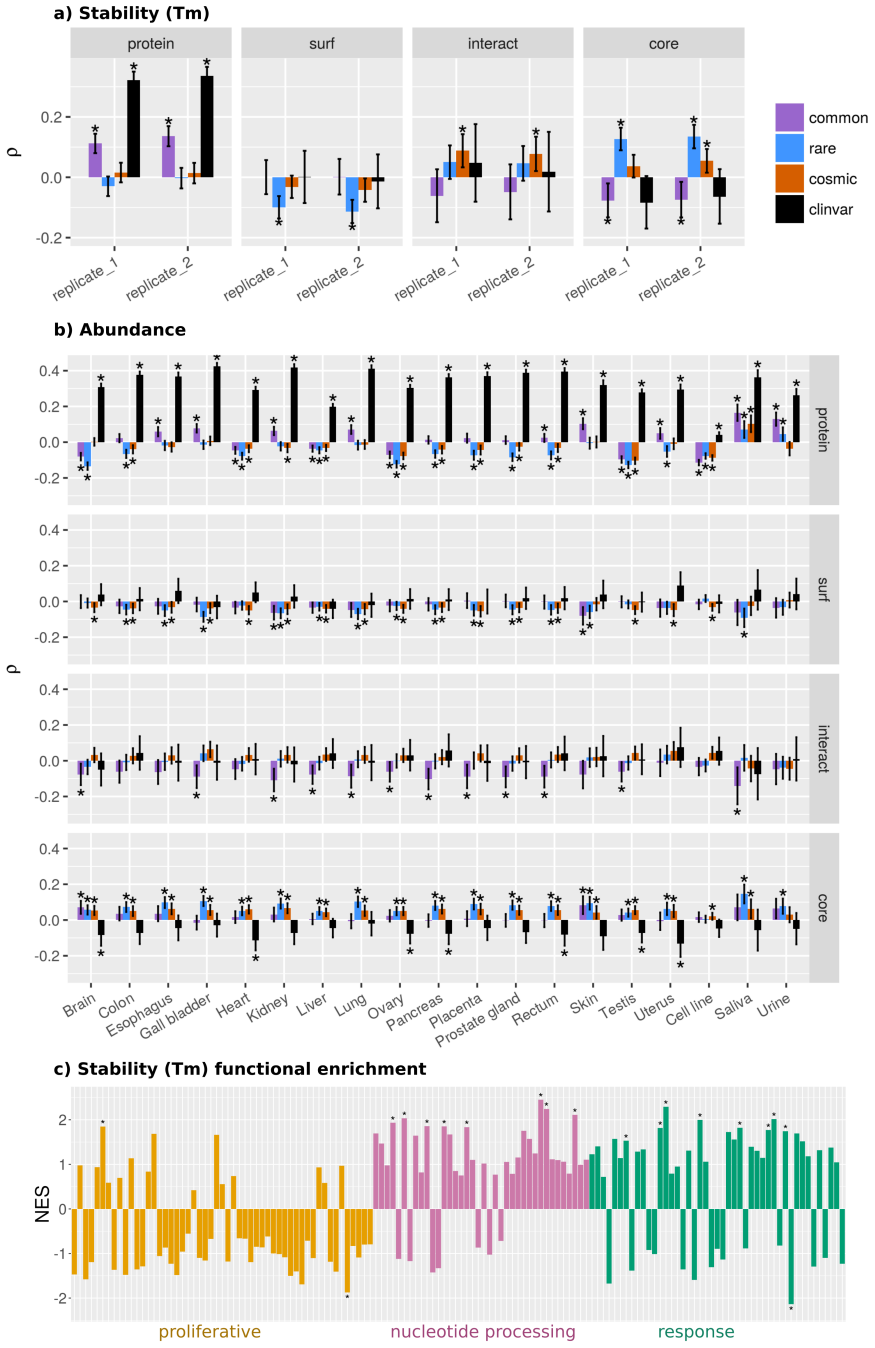

Fig. 3.12 The protein-wise enrichment of SAVs in comparison to protein abundance, expression and stability. Spearman correlations for SAV enrichment (quantified as VESs) at different levels with a) protein melting temperature (Tm) measured in 2 replicates and b) protein abundance (ppm) measured in different tissues/sample-types. The VES is calculated to describe the variant enrichment of a protein in comparison to the entire proteome, and the variant enrichment of a protein region (surf, interact or core) in comparison to the rest of a protein. This is calculated for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). Error bars indicate 95 % confidence intervals. * indicates q-value $<$ 0.05. c) Functional enrichment of proteins in KEGG pathways according to Tm calculated using the GSEA algorithm. The Normalised Enrichment Score (NES) is shown on the y-axis. Pathways have been mapped to the 3 clusters defined in Section 3.3.1 and have been named to reflect their pathway composition.

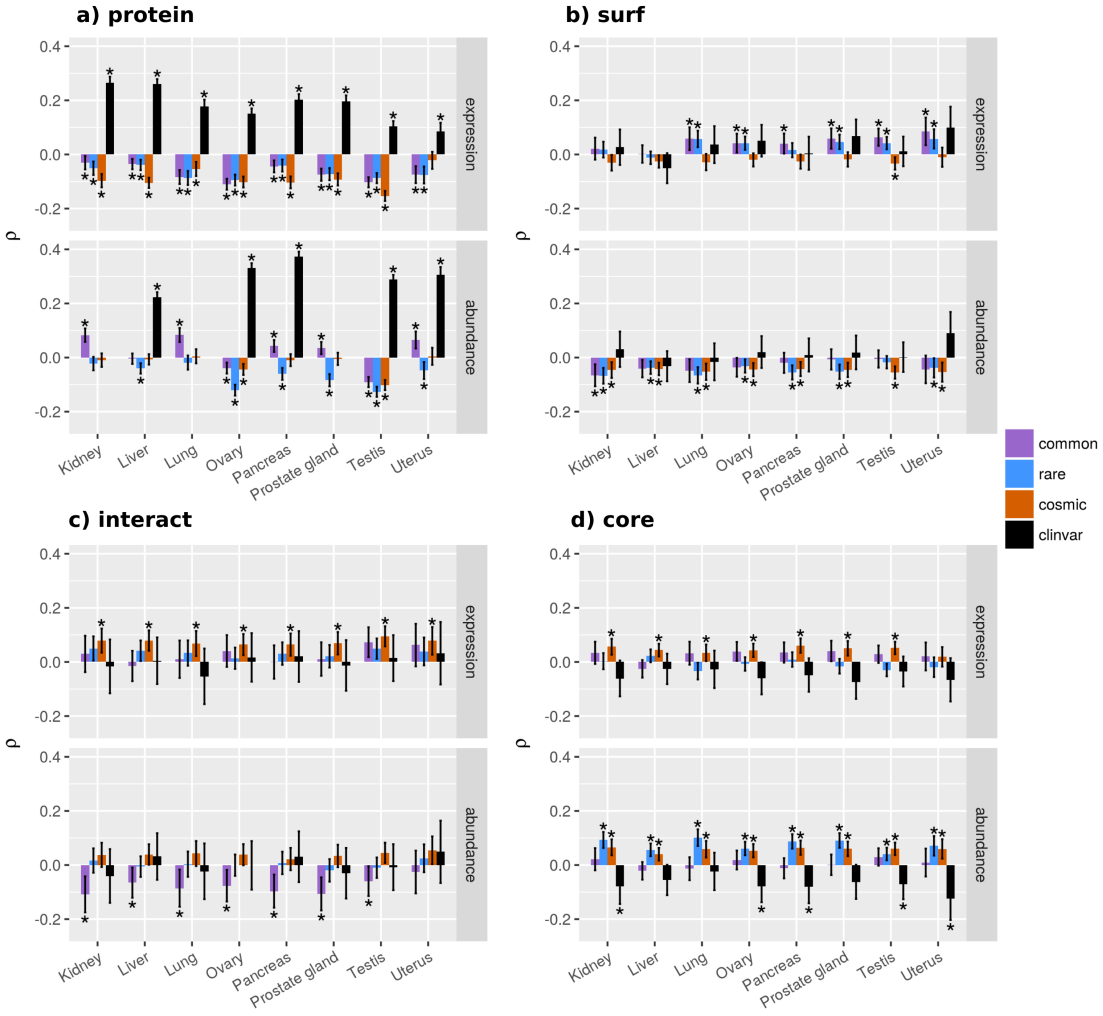

Fig. 3.13 The protein-wise enrichment of SAVs (VES) in comparison to protein abundance (ppm) and expression (median count), measured in different tissues, at the a) "protein" level, b-d) "surf","core" and "interact" levels. The VES is calculated to describe the variant enrichment of a protein in comparison to the entire proteome, and the variant enrichment of a protein region (surf, interact or core) in comparison to the rest of a protein. This is calculated for common population variants, rare population variants, somatic cancer variants (cosmic) and disease-associated variants (clinvar). Spearman correlations calculated using only those proteins present in both abundance and expression data. Error bars indicate 95 % confidence intervals. * indicates q-value < 0.05.

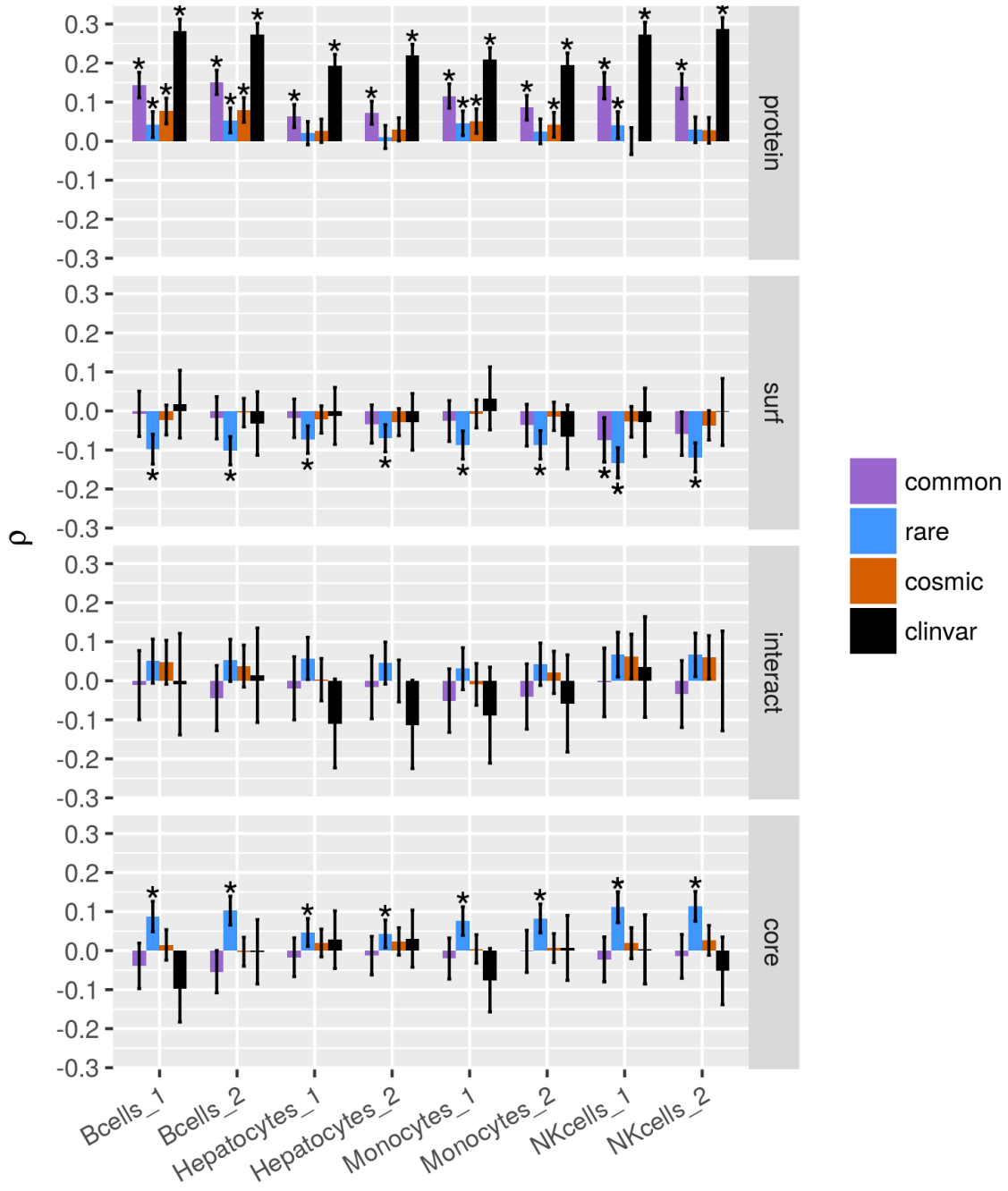

Fig. 3.14 The Spearman correlation of the enrichment of SAVs (VES) with protein half-life data (hours), measured in different cell-types (2 replicas each). The VES is calculated to describe the variant enrichment of a protein in comparison to the entire proteome, and the variant enrichment of a protein region (core, interact or surf) in comparison to the rest of a protein. This is calculated for common population variants, rare population variants, somatic cancer variants (cosmic) and disease-associated variants (clinvar). Error bars indicate 95 % confidence intervals. * indicates q-value < 0.05.

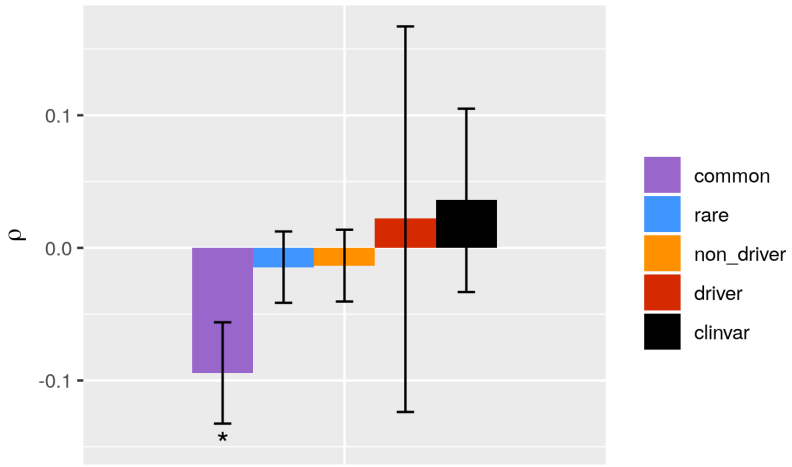

Fig. 3.15 The Spearman correlation of the enrichment of SAVs in protein cores (VES) with a proxy for protein core density (see Section 3.2.4). The VES has been calculated for common and rare population variants, somatic cancer variants (cosmic) and disease-associated variants (clinvar). Error bars indicate 95 % confidence intervals. * indicates q-value < 0.05.

3.3.6 Rare variants are similar to common variants

Throughout the majority of analyses, performed both at the macroscopic and atomistic levels, the greatest segregation of data can be seen between common and disease-associated variants. Rare variants show characteristics more similar to common variants, both in terms of the functional pathways they impact on, and in terms of the protein regions they localise to (core, surface and interface, order and disorder). If more stringent minor allele frequency (MAF) thresholds are used to define rare variants, their properties move towards those of disease-associated variants, but still remain closest to those of common variants (see Figs 3.16-3.17). A visible separation between common and rare variants, especially in the pathway analysis, can only be seen if an extreme MAF cutoff (<0.00001) is used.

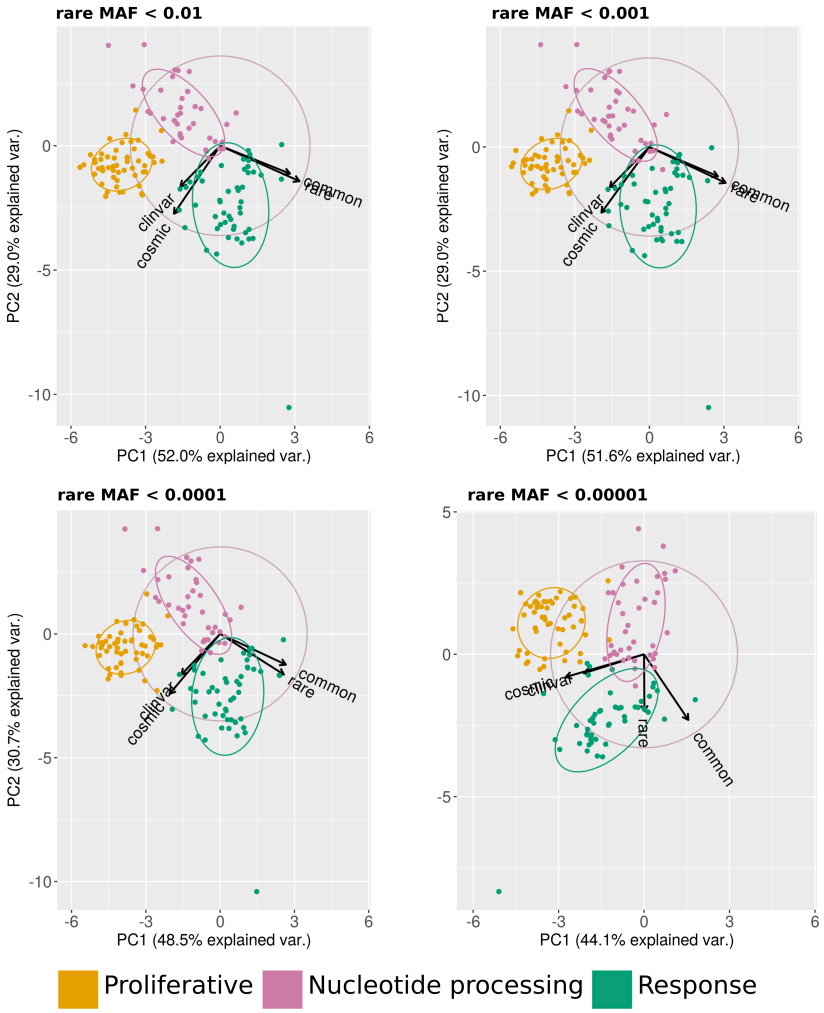

Fig. 3.16 Rare variants display similar functional enrichments to common variants. Gene set enrichment for KEGG functional pathways was performed by ranking proteins using their whole protein VESs, which quantify the variant enrichment of a protein in comparison to the whole proteome. The analysis was performed separately for common population variants (common), rare population variants (rare), somatic cancer variants (cosmic) and disease-associated variants (clinvar). As in Fig. 3.3a, at the whole protein level, KEGG pathways form 3 identifiable clusters (K-means), as visualised projected on the first two principal components of the PCA. Each cluster has been assigned a colour for visualisation purposes, and has been named to reflect its pathway composition. The plots show that if increasingly stringent minor allele frequencies (MAFs) are used to define rare variants, the functional enrichment of the rare dataset becomes slightly more similar to that of the disease-associated datasets (clinvar and cosmic), but remains closest to that of the common dataset, as revealed by factor loadings.

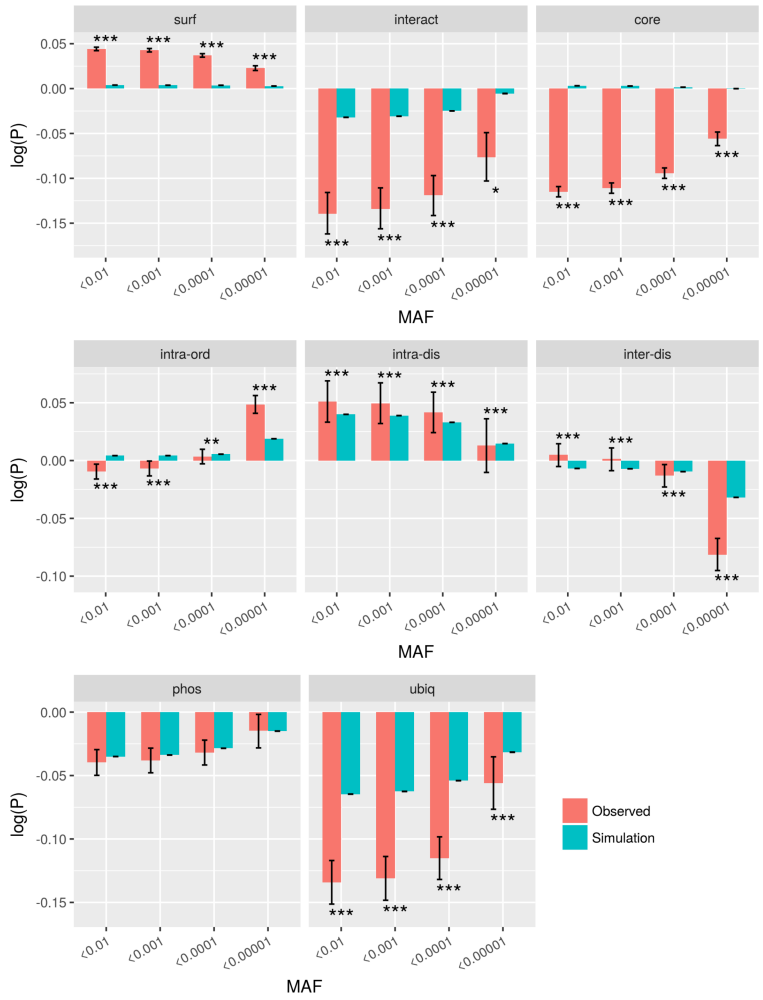

Fig. 3.17 The density of rare mutations from the gnomAD data in different protein regions. Rare variants have been defined using different MAF cut-offs, as shown on the x-axis. Both observed densities (pink), and densities derived from simulated null distributions (turquoise) are shown. Error bars depict 95 % confidence intervals; for observed densities, these were obtained by bootstrapping. Significance was calculated by comparison of observed values to simulated null SAV distributions (significance level indicated by: * q-value $<$ 0.05, ** q-value $<$ 0.001, *** q-value $<$ 0.0001).

3.4 Discussion and conclusions

Throughout this work, we show that SAVs in the general population, considered 'nominally healthy', show properties distinct from those in disease cohorts, both at the macroscopic and atomistic levels. Additionally, although we uncover a spectrum in the properties of variants, which ranges from common population variants to disease-associated ClinVar variants, we find that the properties

of rare variants remain close to those of common variants. These findings contrast with other observations (Mahlich et al., 2017), which suggest that common variants have more impact on molecular function than rare variants. Common variants appear closer in character to disease-associated variants than to rare variants, only for certain proteomics properties, such as the thermal stability and abundance of the variant enriched proteins, as discussed in Section 3.3.5. However, we consider these results inconclusive, due to the sparsity of the data. Alhuzimi et al. (2018) suggest that the properties of genes enriched in rare population variants are similar to those enriched in disease-associated variants, and are thus good candidates for harbouring unknown disease associations. Instead, we show that such proteins are, from the annotated functional pathways, most similar to those enriched in common variants (Fig. 3.16). Moreover our results, which show that variants maintained within a population impact on functions which are mainly associated with response to the environment, such as sensory and immune-related functions (Figs 3.3a and 3.4), agree with results from evolutionary studies reviewed by Quintana-Murci (2016), which used (primarily) a genome-wide scanning approach to detect signals of positive selection.

We have dissected the levels of variant enrichment in diverse datasets and across different protein levels (Fig. 3.2). Such a detailed anatomy of variant enrichment in health and disease provides a unique link between the cataloguing of mutations, and understanding both their mechanistic and functional effects. This supplies invaluable information to researchers studying specific proteins or domains, or focusing on proteins involved in a particular function (e.g. DNA binding; Fig. 3.10). By analysing the enrichment of variants in protein regions (core, surface, interface, disorder and disorder, PTM vicinity), we recapitulate trends observed by previous studies (e.g. in the comparison of oncogenes and TSGs; Fig. 3.6b) (David et al., 2012; Engin et al., 2016; Gao et al., 2015; Gress et al., 2017; Stehr et al., 2011), but also shed light on the debate as to whether somatic cancer variants are enriched in interface regions, by simulating null-distributions of variants. The simulations we have performed show that it is essential to consider that variants from different datasets are not uniformly randomly distributed throughout the proteome. Through density-based metrics we find somatic cancer variants are not enriched in protein interfaces, however using a simulation-based approach we do find an enrichment (Fig. 3.6a). A similar simulation-based approach was taken by Gress et al. (2017), but they found no significant enrichment for COSMIC variants in interface

regions. Whilst they analysed a filtered set of mutations likely to play a driver role, we investigated all somatic variants and addressed separately mutations that localise to defined driver and non-driver genes. Our enrichment calculations were rigorous, and directly compared against null (n = 10,000) simulations to assess statistical significance.

Throughout this analysis, we have, of course, been limited by the number of proteins with available structural data, although this has been enriched by considering homologous structures. Our analyses have attempted to overcome biases which could result from the increased study (and associated structural coverage) of disease-associated proteins, however we cannot fully anticipate the structural properties of the unresolved portion of the proteome. It is likely that this is enriched in proteins with higher flexibility/more disordered regions, which are more difficult to resolve structurally. Therefore, it is possible that greater insight into the properties of missense variants will be achieved as the structural coverage of the proteome increases. However, as the properties we observe are in line with those from earlier studies which had access to a much smaller number of protein structures and variants (e.g. that by Sunyaev et al. (2000)), we consider it unlikely increased structural coverage of the proteome will drastically alter our results, unless techniques for resolving proteins which contain a large portion of disordered/flexible regions continue to improve (Gibbs and Kriwacki, 2018). We are also still limited by the structural coverage of protein interactions; although enough data exists to uncover broad trends, our analyses at a finer granularity, which probed protein-protein interaction sites, generally lacked statistical power. Moreover, it is likely that a more detailed picture will emerge if variant localisations to proteins involved in different classes of interactions are probed (e.g. transient vs permanent interactions). We envisage that the recent advances in cryo-EM (Orlov et al., 2017), and the integration of structural data derived by a variety of techniques (Burley et al., 2017), will further increase the structural coverage of the protein-protein interaction network, enabling such finer-grained analyses in the future.

Our analysis at the macromolecular level, which probes associations between the enrichment of variants and proteomic features, is, to the best of our knowledge, unprecedented, and has only been made possible due to the recent release of large-scale proteomics data (Franken et al., 2015; Leuenberger et al., 2017; Mathieson et al., 2018; Wang et al., 2015). We observe correlations which suggest an interplay between variant enrichment, protein abundance and thermal stability.

First, disease-associated variants localise preferentially to proteins which are highly expressed and abundant (Fig. 3.12). These results complement a body of research which concludes that the rate of protein evolution correlates negatively with protein expression and abundance (Zhang and Yang, 2015). The extent of this anti-correlation has been found to be tissue-specific; those tissues with a high neuron density demonstrating the highest anti-correlation (Drummond and Wilke, 2008). Consistent with this, we found the largest negative correlation for the protein-wise enrichment of rare variants, from the gnomAD dataset, with protein abundance in the brain, and, interestingly also in the ovary and testis, which both harbour long-lived germline progenitor cells (Fig. 3.12b; Fig. 3.13); purportedly the lifespan of long-lived cells renders them more sensitive to the toxicity of misfolded proteins. Second, we see a trend which suggests disease-associated variants preferentially localise to the core in less thermally stable proteins, most probably as these are more easily destabilised to an extent at which function is lost or impaired (Fig. 3.12a). Hence two competing trends emerge; variants which localise to less abundant proteins have greater disruptive potential, conversely, those which localise to thermally unstable proteins (which are normally less abundant (Leuenberger et al., 2017)) may be able to deleteriously destabilise such proteins more easily. It is conceivable that the chemical nature of the particular missense variant plays an important role here: e.g. if a variant at the protein surface alters the "stickiness" of the protein and promotes non-specific interactions, this is likely to be most detrimental if the affected protein is present in great abundance. This highlights the importance of evaluating the interplay of macroscopic and atomistic features when estimating the potential impact of variants on protein function and stability.

The relationship between variant localisation and protein stability is of importance, as a number of algorithms have used the change in protein stability upon mutation ($\Delta\Delta$G) as a proxy for variant impact. Our results indicate that the baseline stability of the wild-type protein may also be important when considering the phenotypic relevance of a change in stability upon mutation. From their analysis of the ProTherm database, Serohijos et al. (2012) found that mutations in more stable proteins generally led to greater destabilisation. They interpret this as suggesting that proteins which have evolved to become more stable are in a state closer to their peak stability, where any changes will result in drastic destabilisation. Similarly, Pucci and Rooman (2016) used temperature dependent statistical potentials to investigate the thermal stability of the structurome (all proteins

with resolved structure), and concluded that mutations in proteins which are highly thermally stable lead to a larger decrease in thermal stability, compared with those in less thermally stable proteins. We believe that our results point to the fact that, even under a scenario in which mutations in proteins with higher stability result in a greater change in stability, a mutation in an already unstable protein is more likely to result in complete/partial unfolding under physiological conditions. These factors should be brought into consideration when interpreting the impact of missense variants.

By investigating the interplay of macroscopic and atomistic features, such as pathway and region enrichment, we show that greater insight into the properties of variants in health and disease can be obtained. For example, as discussed in Section 3.3.1, it can be clearly seen that population variants are most enriched on the surface of proteins which take part in pathways we have defined as belonging to the "proliferative" cluster (Fig. 3.3d). Moreover, pathways belonging to this cluster also appear to be enriched in proteins with less thermal stability (Fig. 3.12c), suggesting a possible mechanistic basis underlying the localisation of variants (variants tend to localise to the surface and avoid disrupting the core of these already unstable proteins). This indicates that the combinatorial use of such features may aid in both improving the prediction of a variant's impact on phenotype, and in assessing the molecular mechanisms underlying this.

Ultimately, the goal should reach beyond the identification of variants which underlie a disease phenotype, to the use of this information to inform the development of therapeutic strategies. Here we envisage that our domain-centric landscape of variant enrichment (Fig. 3.11), which includes the mapping of domain-targeting drugs, besides providing another feature for the characterisation of variants, will allow for more informed decisions in selecting new therapeutic targets. This will allow for the selection of domains to which multiple disease-associated variants localise, in order to give scope for drug repurposing or redesign. Additionally, targets with few population variants could be selected, to minimise differential drug response due to genetic differences between individuals. It is, of course, likely that such a differential drug response may not only be associated with genetic variants which localise to the target protein, but also a number of other factors, such as variants which localise to interacting proteins and environmental conditions. Nevertheless, our analysis provides a starting point for determining the actionability of disease-associated variants and domains. As with all other analyses presented here it must be considered that our knowledge of protein-drug

and domain-drug interactions is incomplete, and depends partially on the techniques which have been used to study a drug, and the particular drug design and development processes involved. For example, the mechanisms of action (and target binding) for a number of drugs, including paracetamol (Mehta and Sharma, 2013), are unknown. It is possible that systematic biases could result from this; for example drug-target interactions with a lower binding affinity may be less frequently characterised.

In conclusion, this chapter highlights the complex interplay between different factors which may determine variant pathogenicity, at both the macroscopic and atomistic levels. We believe that these insights will prove important in the prediction of which variants drive disease phenotypes. Further advancement in the structural coverage of the proteome, and the exploitation of high throughput proteomics technologies, such as those pioneered by the Savitski and Picotti labs (Leuenberger et al., 2017; Mathieson et al., 2018), will offer a finer-grained picture of features which segregate variants in "health" and "disease".

Chapter 4

Predicting the impact of titin variants using protein dynamics-based features

4.1 Introduction

As discussed in Chapter 3, despite the explosion of accessible genetic data the problem of missing heritability still persists. A number of variants defy classification by purely statistical methods, therefore their impact must be elucidated by a combination of computational and wet lab techniques. Furthermore, the impact of a missense variant on structure and function must be discerned, if the mechanism underlying a variant's pathogenicity is to be understood. Such insights may inform therapeutic strategies. Although a number of computational methods have been developed to assess the impact of variants, it is clear these do not always match experimental outcomes. Moreover, as discussed in Chapter 1, evidence suggests that existing methods lack specificity.

One important aspect of proteins' functional roles is their dynamic behaviour; in fact proteins are not static by nature, but in a perpetual state of motion within the cell. Therefore the impact of a variant may act on particular conformational states and transitions between these. However, we have seen in Chapter 1 that very few predictors use features associated with protein flexibility or dynamics. Importantly, recent work from the Bahar laboratory has shown that incorporating structural dynamics information derived from elastic network models can improve the accuracy of prediction (Ponzoni and Bahar, 2018). However, as the dynamic features within this approach

are based solely on Cα network models, the chemical nature of the change is not modelled. Thus this dynamic information only allows discrimination between positions which are most likely to harbour deleterious mutations. Atomistic molecular dynamics, as discussed in Section 1.4.2, may enable more accurate modelling of the physicochemical nature underlying a residue's flexibility, and therefore allows for a more realistic representation of the impact of a variant on protein dynamics. Yet, to date, atomistic molecular dynamics simulations have generally only been used to investigate a small number of "case study" variants, or been applied over very short timescales to refine modelled mutants (see Section 1.4.2). Here, an obstacle to the large-scale use of such simulations is their computational cost. Because of this the use of coarse-grained methods, which contain approximate but nonetheless realistic representations of amino acids, and are less computationally expensive, would be desirable. However, it is not clear whether these are able to reflect the impact of mutations on dynamics.

As the focus of this thesis is on titinopathies, we explore here whether population titin variants, the majority of which are rare, can be distinguished from titin variants with known disease associations. In particular, we investigate the application of dynamics features to the task of predicting the impact of these variants. We calculate elastic network features, for titin Ig and Fn3 domains, as in the work by Ponzoni and Bahar (2018) however we go further to explore the use of coarse-grained molecular dynamics using the CafeMol (Kenzaki et al., 2011), Martini ElNeDyn (Periole et al., 2009) and Martini GO models (Poma et al., 2017), in addition to full atomistic molecular dynamics in explicit solvent. Each of these models can be placed upon a spectrum which ranges from the Cα representation (which explores only near-equilibrium dynamics) of elastic network models, to the Cα representation with residue specific potentials of the CafeMol simulations (Kenzaki et al., 2011), to the Martini representation in which groups of atoms are represented by virtual "beads" (de Jong et al., 2013; Monticelli et al., 2008), and finally to full atomistic molecular dynamics simulations (Meier et al., 2013; van Gunsteren et al., 2006, 2018)(see Fig. 4.1).

We ask whether atomistic molecular dynamics simulations are better able to distinguish between disease-associated and (predominantly) rare population variants than coarse-grained models and elastic network models. Furthermore, we compare the predictive power of features derived from atomistic molecular dynamics simulations to those derived from ENMs and sequence-based features,

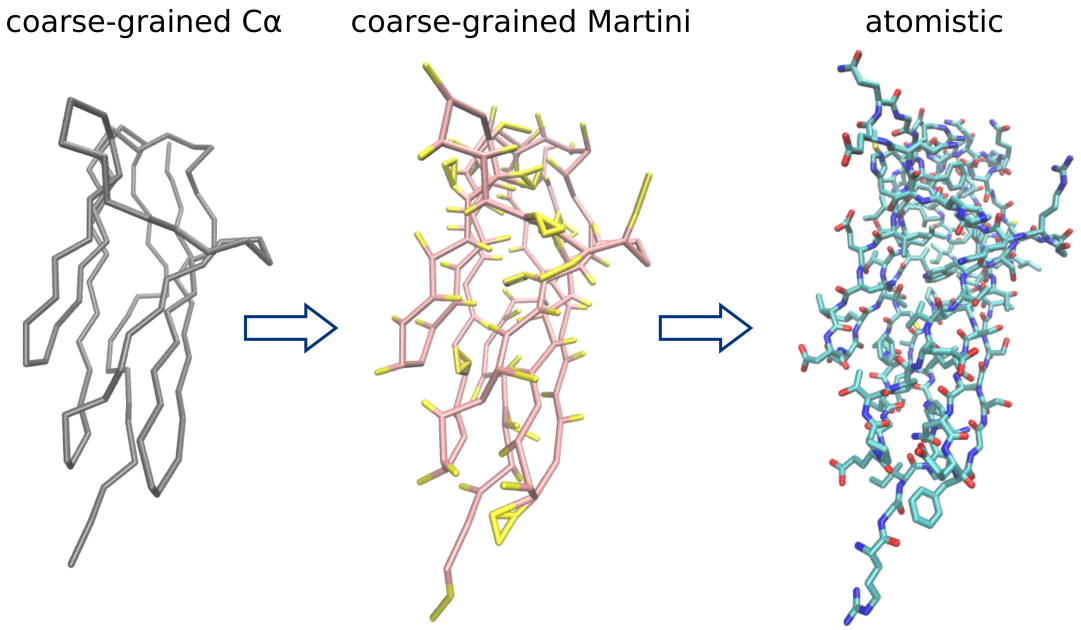

Fig. 4.1 Different levels of granularity can be used to represent protein structures. These include, in order of increasing detail, coarse-grained Cα representations, coarse-grained Martini representations, in which several atoms are grouped together to form virtual beads, and atomistic representations. PDB structure 2y9r is depicted, as visualised by the VMD software (Humphrey et al., 1996).

by training random forest-based predictive models to classify disease-associated and population variants. Finally, we compare the performance of these models to that of widely used predictors, such as PolyPhen2 (Adzhubei et al., 2010) and FATHMM (Shihab et al., 2013).

Unfortunately, a perfect dataset to test whether rare non-pathogenic variants can be segregated from rare pathogenic variants does not exist. Firstly, there is no guarantee that rare variants may not have an undiscovered disease association. Moreover, disease-associated homozygous and compound heterozygous variants can be present in healthy heterozygotes. The extent to which rare variants may be associated with disease is a matter of active research and debate (Gibson, 2012; Kido et al., 2018). Secondly, we cannot rule out the possibility that variants in our disease-associated set could be incorrectly labelled, due to co-inheritance with a causative allele. However, we believe that this detailed study of the impact of variants on the dynamics of these ubiquitous domains (Ig and Fn3) may shed light on the proportion of rare variants which may have an impact on protein function.

4.2 Materials and methods

4.2.1 Dataset

19 titin variants were selected based on their disease associations and localisation to the constitutively spliced-in A-band and M-line regions of titin. These variants localise to 5 distinct titin domains. For each domain, an equal number of variants with no known disease-associations were selected from the gnomAD database, based on either having a WT solvent accessibility similar to the disease-associated WT positions, or available biophysical data. The rationale being that, as the majority of disease-associated variants have low solvent accessible surface areas (SASAs), surface exposed population variants are easy to distinguish from these, and thus the use of dynamics-related features is not necessary to achieve segregation. From the minor allele frequencies it can be seen that the majority of selected SAVs are extremely rare; only the Fn3-90 I14V has a MAF above the cut-off of 0.01, which is frequently used to define common variants. One additional population variant was selected for the domain Fn3-90, due to the availability of associated in-house biophysical data [2] for two population variants from this domain. Selected deleterious and population variants are detailed in Table 4.1 and Table 4.2. In particular, two of the domains, Ig-169 and Fn3-119, are hotspots for disease-associated mutations. The domain Ig-169, also known as M10, is located at the end of titin's M-band region where it interacts with obscurin and obscurin-like protein. This domain harbours variants associated with tibial muscle dystrophy (TMD) (Hackman et al., 2002; Savarese et al., 2016) and limb-girdle muscular dystrophy 2J (LGMD2J) (Hackman et al., 2002; Savarese et al., 2016). Similarly, the domain Fn3-119 harbours a number of variants associated with the disease hereditary myopathy with early respiratory failure (HMERF) (Pfeffer et al., 2015). For all investigated domains, variants are depicted mapped to structure in Fig. 4.2.

[2] Private communication, Roksana Nikoopour and Dr Martin Rees (Gautel laboratory)

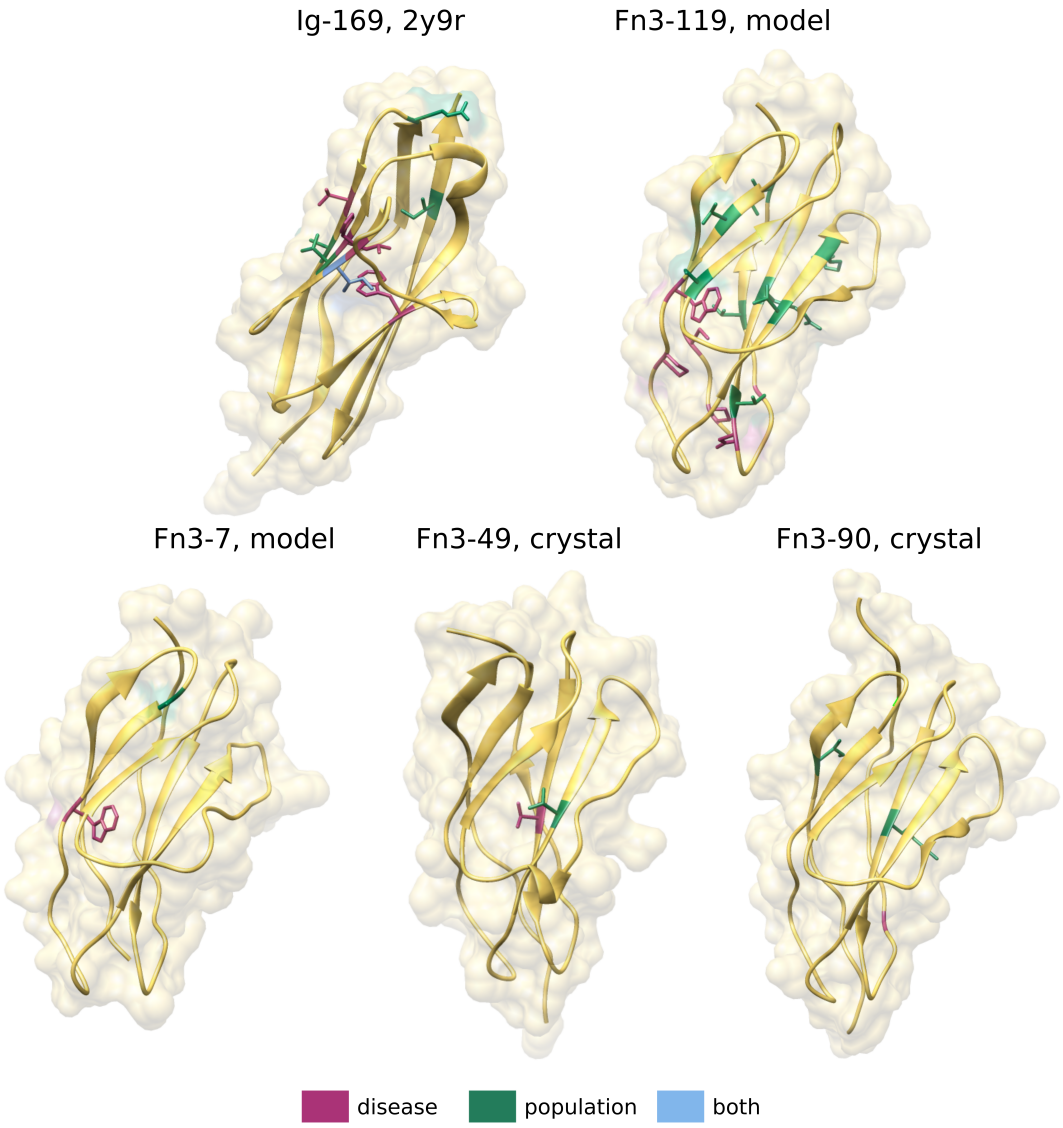

Fig. 4.2 The location of population variants (green) and disease-associated variants (magenta) analysed in this study, mapped to structures for the titin domains Ig-169, Fn3-119, Fn3-7, Fn3-49 and Fn3-90. One position (blue), in Ig-169, harbours distinct variants from the disease-associated and population datasets. Structures were visualised using the UCSF Chimera software (Pettersen et al., 2004).

Domain name	SAV	Position (IC)	Q(SASA)	Phenotype	Source	Structure	GnomAD MAF	WT Tm	Mutant Tm
Fn3-7	W22C	17471	0.10	M-DCM	in-house	model	4.00E-04	75*	43*
Fn3-49	V81E (V77E)	22232	0.09	MmD-HD	literature	in-house		59*	unfolded
Fn3-90	G88V (G84V)	27849	0.17	M-DCM	in-house	in-house		60*	unfolded*
Fn3-119	P2R	31709	0.15	HMERF	literature	model		50*	
Fn3-119	C5Y	31712	0.12	HMERF	literature	model		50*	
Fn3-119	C5R	31712	0.12	HMERF	literature	model	4.164E-06	50*	32*
Fn3-119	W22R	31729	0.11	HMERF	literature	model		50*	
Fn3-119	W22L	31729	0.11	HMERF	literature	model		50*	
Fn3-119	W22C	31729	0.11	HMERF	literature	model		50*	
Fn3-119	P25L	31732	0.15	HMERF	literature	model	1.22E-05	50*	33*
Fn3-119	N79K	31786	0.12	HMERF	literature	model		50*	
Fn3-119	G84V	31791	0.15	HMERF	literature	model		50*	
Fn3-119	G84R	31791	0.15	HMERF	literature	model		50*	
Fn3-119	G84D	31791	0.15	HMERF	literature	model		50*	
Ig-169	H54P (H50P)	35946	0.21	TMD	literature	2y9r		58	37
Ig-169	I55N (I51N)	35947	0.14	TMD	literature	2y9r	4.06E-06	58	53
Ig-169	L64P (L60P)	35956	0.11	TMD	literature	2y9r		58	unfolded
Ig-169	T23P (T19P)	35915	0.32	TMD	literature	2y9r		58	
Ig-169	W38R (W34R)	35930	0.15	LJMD2J	literature	2y9r		58	

Table 4.1 Titinopathy associated variants analysed in this study. Note that a minority of these are present in the gnomAD database at low MAFs. SAVs are numbered by their position in the domain structure. TITINdb domain position numbering (Laddach et al., 2017) is included in brackets where this does not match the numbering of the structure used here. In-house differential scanning fluorimetry (DSF) data [2] is denoted by *. All other Tm measurments are from circular dichromism (CD) data reported in (Rudloff et al., 2015). Titinopathies associated with these variants include hereditary myopathy with early respiratory (HMERF), tibial muscular dystrophy (TMD), limb-girdle muscular dystrophy 2J (LGMD2J), multi-minicore disease with heart disease (MmD-HD) and myopathy with dilated cardiomyopathy (M-DCM).

Domain name	SAV	Position (IC)	Q(SASA)	GnomAD MAF	1000 genomes MAF	Structure	WT Tm	Mutant Tm
Fn3-7	A17T	16466	0.10	8.23E-06		model	75*	
Fn3-49	V42I (V38I)	31748	0.09	8.20E-06		in-house	59*	
Fn3-90	I14V (I10V)	27775	0.39	3.52E-1	5.84E-1	in-house	60*	59*
Fn3-90	R78Q (R74Q)	27839	0.17	4.21E-04	3.99E-04	in-house	60*	47*
Fn3-119	T19A	31726	0.16	4.08E-06		model	50*	
Fn3-119	V38M	31745	0.09	3.26E-05		model	50*	
Fn3-119	R41H	31748	0.10	1.22E-05		model	50*	
Fn3-119	R41C	31748	0.10	1.22E-05	1.01E-3	model	50*	
Fn3-119	R41S	31748	0.10	4.07E-06		model	50*	
Fn3-119	S59F	31766	0.33	3.39E-3	5.99E-3	model	50*	49*
Fn3-119	R74L	31781	0.12	8.15E-06		model	50*	
Fn3-119	R74Q	31781	0.12	1.63E-05		model	50*	
Fn3-119	V78I	31785	0.14	4.07E-06		model	50*	
Fn3-119	V87L	31799	0.14	3.23E-05		model	50*	
Fn3-119	I92V	31799	0.16	6.12E-05		model	50*	
Ig-169	E18K (E14K)	35910	0.33	1.23E-05	1.20E-4	2y9r	58	
Ig-169	A25V (A21V)	35917	0.16	4.06E-06		2y9r	58	
Ig-169	I55T (I51T)	35947	0.14	4.08E-06		2y9r	58	
Ig-169	T63N (T59N)	35955	0.21	4.08E-06		2y9r	58	
Ig-169	I94T (I90T)	35986	0.12	3.39E-3		2y9r	58	

Table 4.2 Population variants analysed in this study. SAVs are numbered by their position in the domain structure. TITINdb domain position numbering (Laddach et al., 2017) is included in brackets where this does not match the numbering of the structure used here. In-house differential scanning fluorimetry (DSF) data [2] is denoted by *. All other Tm measurments are from circular dichromism (CD) data reported in (Rudloff et al., 2015).

4.2.2 Modelling of domains

Homology models of Fn3-119 and Fn-7, were obtained from TITINdb (Laddach et al., 2017). Mutants were created using the Modeller protocol available at http://salilab.org/modeller/wiki/Mutate%20model (Webb and Sali, 2014).

4.2.3 Molecular dynamics

Molecular dynamics is an established method which enables the simulation of the dynamical properties of a system over time, using Newton's equations of motion (Hospital et al., 2015).

The force f_i on particle i, with mass m_i, is related to its acceleration:

$$f_i = m_i \frac{\delta v_i}{\delta t_i} \tag{4.1}$$

The velocity v of particle i after time τ under constant force, given an initial velocity $v_i(0)$, is described by the equation:

$$v_i(\tau) = v_i(0) + \int_0^\tau \frac{dv_i}{dt} dt \tag{4.2}$$

And the position r of the particle is given by the equation:

$$r_i(\tau) = r_i(0) + \int_0^\tau v_i(t) dt \tag{4.3}$$

However, once a particle moves, the forces acting on it change. Therefore, in the context of simulations, velocities, forces and positions are updated at finite time steps of chosen length Δt. Several algorithms can be used to calculate forces and velocities at subsequent times, according

to discretisation choices. The molecular dynamics engine GROMACS, used for both atomistic and Martini simulations, employs the leapfrog algorithm:

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}F(t) \tag{4.4}$$

$$r(t + \Delta t) = r(t) + \Delta t v(t + \frac{1}{2}\Delta t) \tag{4.5}$$

This is used for both the atomistic molecular dynamics simulations and the coarse-grained Martini simulations. For CafeMol simulations the velocity Verlet algorithm is instead used to update velocities, forces and positions:

$$r(t + \Delta t) = r(t) + \Delta t v + \frac{\Delta t^2}{2m}F(t) \tag{4.6}$$

$$v(t + \Delta t) = v(t) + \frac{\Delta t}{2m}[F(t) + F(t + \Delta t)] \tag{4.7}$$

This algorithm is similar to the leapfrog algorithm, although velocities and positions are both calculated at the same time. Because of this, velocities must be explicitly calculated, however the algorithm has the advantage that it is self-starting, i.e. it is possible to calculate velocities and positions at time $t + \Delta t$ from those at time t. In comparison, the leapfrog algorithm requires information from the previous time step.

The calculation of force acting on each particle of the system is accomplished using molecular mechanics force fields. Forces are calculated from the ensemble of all particle positions, denoted as r, within the system.

For all atom simulations we use the Gromos 54a8 forcefield (Reif et al., 2013). This consists of the following terms, which use the ensemble of all positions to account for bonded interactions (bond stretching, bond-angle bending, improper and proper dihedrals) (van Gunsteren et al., 2006):

$$V^{bond}(\mathbf{r}; K_b, b_0) = \sum_{n=1}^{N_b} \frac{1}{4} K_{b_n} [b_n^2 - b_{0_n}^2]^2 \tag{4.8}$$

$$V^{angle}(\mathbf{r}; K_\theta, \theta_0) = \sum_{n=1}^{N_\theta} \frac{1}{2} K_{\theta_n} [cos(\theta_n) - cos(\theta_{0_n})]^2 \tag{4.9}$$

$$V^{improper}(\mathbf{r}; K_\xi, \xi_0) = \sum_{n=1}^{N_\xi} \frac{1}{2} K_{\xi_n} [\xi_n - \xi_{0_n}]^2 \tag{4.10}$$

$$V^{proper}(\mathbf{r}; K_\varphi, \delta, m) = \sum_{n=1}^{N_\varphi} K_{\varphi_n} \left[1 + cos(\delta_n)\, cos(m_n \varphi_n)\right] \tag{4.11}$$

Here K_b, b_0 and b_n represent the bond force constant, the optimum bond length and the observed bond length. K_θ, θ_0 and θ_n represent the angle force constant, optimum bond angle and observed bond angle. K_ξ, ξ_0 and ξ_n represent the improper dihedral force constant, optimum improper dihedral and observed improper dihedral. It is important to note that improper dihedrals play an important role in maintaining chirality. Finally, K_φ, δ, m and φ_n represent the proper dihedral force constant, the phase shift, the multiplicity and the observed proper dihedral angle.

Additionally two terms are used to account for non-bonded interactions (van der Waals, and electrostatic interactions) (see equations 4.12 and 4.13). For all atomistic and Martini simulations performed here, the Lennard-Jones potential was used to model van der Waals interactions.

$$V^{LJ}(\mathbf{r}; C_{12}, C_6) = \sum_{pairs\, i,j} \left(\frac{C_{12}(i,j)}{r_{i,j}^{12}} - \frac{C_6(i,j)}{r_{i,j}^{6}} \right) \tag{4.12}$$

$$V^C(\mathbf{r};q) = \sum_{pairs\, i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1} \cdot \frac{1}{r_{i,j}} \tag{4.13}$$

Here C_{12} and C_6 are parameters of the 12/6 Lennard Jones interaction. The parameters q_i and q_j denote the partial charges on atoms. Long-range electrostatics must also be accounted for. In this work we use the Particle-mesh Ewald method (Essmann et al., 1995) for atomistic simulations and the reaction field method (Tironi et al., 1995) for Martini coarse-grained simulations.

The Martini coarse-grained forcefield groups together, on average 4 heavy atoms, to form virtual beads. Four different bead types are possible; polar, nonpolar, apolar and charged (Monticelli et al., 2008). Each bead type is further split into subtypes based on polarity and hydrogen bonding capacity. Forcefield terms are identical to those used by the Gromos forcefield. Here it is important to note that, in the coarse-grained representation, proper dihedrals are used to enforce secondary structure, thereby preventing secondary structure transitions.

The CafeMol software represents protein Cα atoms as beads, however uses residue specific potentials and all atom information to define local contacts (Takada et al., 2015). The CafeMol AICG2+ force field contains the following local force-field terms:

$$V^{bond}(\mathbf{r}|\mathbf{r_0};K_{bond}) = \sum_{n=1}^{N_b} K_{bond}(b_n^2 - b_{n_0})^2 \tag{4.14}$$

$$V^{pseudo-angle}(\mathbf{r}|\mathbf{r_0};K_{pseudo-angle},W_{pseudo-angle}) = \sum_{j=i+2} K_{pseudo-angle} exp\left(-\frac{(r_{i,j}-r_{i,j_0})^2}{2W^2_{pseudo-angle}}\right) \tag{4.15}$$

$$V^{proper}(\mathbf{r}|\mathbf{r_0};K_\varphi,W_\varphi) = \sum_{n=1}^{N_\varphi} K_\varphi exp\left(-\frac{(\varphi_n-\varphi_{n_0})^2}{2W^2_\varphi}\right) \tag{4.16}$$

Here each term is calculated given properties of the native structure (denoted by subscript zero), with the native structure frequently defined by the initial coordinates. For example b_{n0} denotes the bond length in the native structure. Distances between residue i and $i+2$ are used in a pseudo bond angle term, and this potential is modelled as a Gaussian distribution. The dihedral angle term V^{trig} is also modelled as a Gaussian. For both these terms the strength of the interaction is given by the constants $K_{pseudo-angle}$ and K_{φ}, and the widths of the Gaussians are controlled by the constants $W_{pseudo-angle}$ and W_{φ}. Additionally a flexible local potential, V_{loc}^{flp} is used to account for chirality. To construct this potential, statistical probability distributions for bond angles and dihedral angles were extracted from PDB structures. This gives the following potential energy functions:

$$V_{ba}^{stat} = -k_B T ln \frac{P(\theta)}{sin(\theta)} \tag{4.17}$$

$$V_{dih}^{stat} = -k_B T ln P(\varphi) \tag{4.18}$$

Here k_B denotes the Boltzmann constant. The temperature T is 300 K. To create continuous potential energy functions, cubic spline interpolation is used for bond angle distributions and dihedral distributions are fitted to the truncated Fourier series.

Additionally, the CafeMol forcefield employs the following non-local forcefield terms:

$$V^{natcontact}(\mathbf{r}|\mathbf{r_0}; K_{go}) = \sum_{pairs\, i,j}^{natcontact} K_{go} \left[5 \left(\frac{r_{i,j_0}}{r_{i,j}} \right)^{12} - 6 \left(\frac{r_{i,j_0}}{r_{i,j}} \right)^{10} \right] \tag{4.19}$$

$$V^{non-native}(\mathbf{r}|\mathbf{r_0}; K_{ev}) = \sum_{pairs\, i,j}^{non-native} K_{ev} \left(\frac{d_i + d_j}{2r_{i,j}} \right)^{12} \tag{4.20}$$

Native contacts are non-local pairs which are close to one another in the reference structure. Non-native pairs are non-local pairs which are not close to one another in the reference structure. The parameters d_i and d_j denote residue-type specific radii.

A term to account for hydrophobic interactions takes the form:

$$V^{HP} = -c_{HP} \sum_{i \in HP} K_{HP,A(i)} S_{HP}(\rho_i) \tag{4.21}$$

Where the parameter c_{HP} controls the strength of hydrophobic interactions, $A(i)$ is the particle type (i.e. the amino acid type) and the function S_{HP} calculates the "buriedness" of particle i from the local density ρ_i:

$$S_{HP}(\rho) = \begin{cases} 1, & \text{if } \rho \geq 1 \\ c_{linear}\rho + 0.5(1 - c_{linear}) \left[1 + cos\frac{\pi(1-\rho)}{1-\rho_{min}}\right], & \text{if } \rho_{min} < \rho < 1 \\ c_{linear}\rho, & \text{if } \rho \leq \rho_{min} \end{cases} \tag{4.22}$$

Here c_{linear} is a constant for the calculation of particle buriedness and ρ_{min} is the minimum local density of a particle. The parameter ρ_i is calculated using the following formula:

$$\rho_i = \frac{\sum_{j \in HP, j \neq i} n_{A(j)} u_{HP}(r_{i,j}, r_{min,A(i),A(j)}, r_{max,A(i),A(j)})}{n_{max,A(i)}} \tag{4.23}$$

Where the degree of contact between the particles i and j is represented by the function u_{HP}:

$$u_{HP}(r, r_{min}, r_{max}) = \begin{cases} 1, & \text{if } r \leq r_{min} \\ 0.5\left(1 + cos\frac{\pi(r - r_{min})}{r_{max} - r_{min}}\right), & \text{if } r_{min} < r < r_{max} \\ 0, & \text{if } r \geq r_{max} \end{cases} \tag{4.24}$$

Here n_A is the number of atom types that particle A represents, $n_{max,A}$ is the maximum coordination number for the particle type A, $r_{i,j}$ is the distance between hydrophobic particles i and j, $r_{min,A,B}$ defines the cut-off for the minimal distance between particle types A and B and $r_{max,A,B}$ defines the cut-off for the maximal distance between particle types A and B.

Debye-Hückel type electrostatic interactions are also implemented, to implicitly represent the screening action of the solvent:

$$V^{ele}(\mathbf{r}|\mathbf{r_0};q) = \sum_{pairs\ i,j} \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_1 r_{i,j}} e^{\frac{-r_{i,j}}{\lambda_D}} \tag{4.25}$$

$$\lambda_D = \left(\frac{\varepsilon_0\varepsilon_1 k_B T}{2N_A e^2 I}\right)^{0.5} \tag{4.26}$$

Where λ_D is the Debye-Hückel length, N_A is the Avogadro number and I is the ionic strength.

4.2.4 Atomistic simulation parameters

Atomistic molecular dynamics simulations were performed using the GROMACS 2016.3 package (Van Der Spoel et al., 2005) and the GROMOS 54a8 forcefield (Reif et al., 2013). Each titin domain structure was centred in a triclinical simulation box filled with SPC-E water molecules (Leontyev and Stuchebrukhov, 2010), with the minimal distance between the protein and box boundaries set at 15 Å. An appropriate number of solvent molecules were replaced with sodium or chloride ions to neutralise the system. All bonds were constrained using the LINCS method (Hess, 2008). The system was energy minimised using the steepest descent minimisation algorithm over 2000 steps with positional restraints and 10000 steps without such restraints, both with the time step set at 0.001 ps. Equilibration was carried out first in an NVT ensemble at 50 K, 100 K, 200 K and 300 K respectively, with positional restraints at force constants of 2000, 1000, 500 and 250 kJ/mol nm^2 for 50000 steps each. Subsequently, an NPT ensemble was used at 100 K, 200 K and 300 K for 50000 steps each, with an isotropic coupling at 1 bar pressure and 4.5E-5 bar^{-1} compressibility. A random velocity drawn from the Maxwell-Boltzmann distribution was used at the start of equilibration. Temperature and pressure coupling was achieved using the Berendsen thermostat and barostat (Berendsen et al., 1984).

Every system was simulated for a 100 ns production run, in an NPT ensemble at 1 bar pressure. Here the Parrinello-Rahman algorithm was used for pressure and temperature coupling (Parrinello and Rahman, 1981). For minimisation, equilibration and production runs, the particle mesh Ewald method (Essmann et al., 1995) was used to calculate the long-range electrostatic contribution to the non-bonded interactions with a cut-off of 14 Å, a Fourier spacing of 1.2 Å and cubic interpolation. Similarly a cut-off of 14 Å was used for van der Waals interactions, with a plain cut-off scheme for the long-range treatment of these. The time step for equilibration and production runs was set to 0.002 ps.

4.2.5 Martini simulation parameters

Martini simulations were performed in GROMACS 2018.02 using the Martini-2.2 ElNeDyn (Periole et al., 2009) and GO (Poma et al., 2017) coarse-grained force fields for biomolecules. These approaches were chosen, as preliminary investigations showed that Martini simulations without additional restraints resulted in a loss of a reasonable structural representation even for WT domains (see Fig. 4.3). The ElNeDyn method maintains the structure of a protein using harmonic spring potentials as in elastic network models, whereas the GO approach uses Lennard-Jones potentials based on contacts in the starting structure. The GO approach has the advantage over the ElNeDyn model that it does not rely on harmonic bonds which cannot be broken. Therefore, although still biased by the starting structure, it is likely that more conformational space can be explored if the system is perturbed. Each structure was converted into its coarse-grained representation using the *martinize.py* script (de Jong et al., 2013), and was then solvated with standard Martini water in a cubic box with a minimum distance of 30 Å between the protein beads and the edge of the box. All bonds were constrained using the LINCS method (Hess, 2008). Minimisation in a vacuum was carried out for 10 steps, followed by 5000 steps of minimization with solvent. The steepest descent algorithm was used. Equilibration with restraints, at force constants of 2000, was carried out for 500000 steps (10 ns) in an NPT ensemble, starting with a random velocity drawn from the Maxwell-Boltzmann distribution. Each equilibrated system was simulated for a 100 ns production run. The Parrinello-Rahman algorithm (Parrinello and Rahman, 1981) was used for pressure and temperature coupling. The reaction field method was used for the calculation of long-range electrostatic interactions (Tironi et al., 1995), and

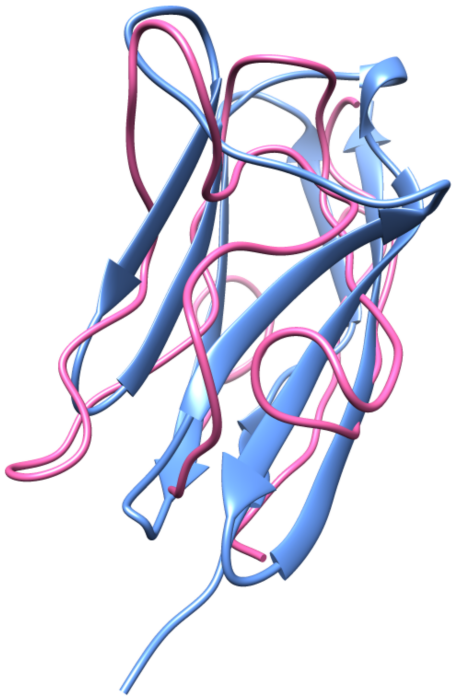

Fig. 4.3 PDB structure 2y9r (domain Ig-169) depicted in blue aligned to its backmapped structure (Wassenaar et al., 2014) after 100 ns simulation using the Martini forcefield (pink). It can be seen that a reasonable representation of protein structure is lost during the simulation. Structures were visualised using the UCSF Chimera software (Pettersen et al., 2004).

a plain cut-off scheme was used for van der Waals interactions. A cut-off of 11 Å was used for both Coulomb and van der Waals interactions. The time step for all runs was set to 0.02 ps, however, caution must be taken in interpreting this time step. As the coarse-graining approach results in a smoother energy landscape, 1 ps in Martini time is roughly equivalent to 4 ps in real time.

4.2.6 CafeMol simulation parameters

Simulations were performed at 300 K, starting with a random velocity drawn from the Maxwell-Boltzmann distribution. The Berendsen thermostat was used for temperature-pressure coupling. Simulations were performed for 2.4082E7 steps, with a time step of 0.1 cafe-time. 1 cafe time $\approx$ 49 fs, however, it is noted that caution must be taken in interpreting this timescale as intrinsic dynamics

are accelerated by the coarse-graining approach. The Debye length was set to 20 Å (this controls ionic screening), the ionic strength to 0.05 mol L^{-1} and the dielectric constant to 78.

4.2.7 Elastic network models

Anisotropic and Gaussian elastic network models were calculated using the ProDy software (Bakan et al., 2011), using default parameters. In both Gaussian and Anisotropic network models, protein Cαs are modelled as nodes. If nodes are within a particular distance cutoff R_c they are connected by springs (Lezon et al., 2010). R_c is set to 10 Å for GNMs and 15 Å for ANMs (ProDy software default parameters).

The GNM potential energy function is based on Cα distances as follows:

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}(\Delta r_j - \Delta r_i)^2\right] = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Delta r_i \Gamma_{ij} \Delta r_j\right] \tag{4.27}$$

Where γ is the spring force constant and Γ is the Kirchoff matrix, which is defined by inter-residue contacts:

$$\Gamma = \begin{cases} -1, & \text{if } i \neq j \ \& \ r_{ij} \leq r_c \\ 0, & \text{if } i \neq j \ \& \ r_{ij} > r_c \\ -\sum_{j,j\neq i}^{N}\Gamma_{ij}, & \text{if } i = j \end{cases} \tag{4.28}$$

This can be written according the X, Y and Z components of Cα fluctuations:

$$V_{GNM} = \frac{\gamma}{2}[\Delta X^T \Gamma \Delta X + \Delta Y^T \Gamma \Delta Y + \Delta Z^T \Gamma \Delta Z] \tag{4.29}$$

The probability distribution for fluctuations is modelled as a Gaussian isotropic distribution.

$$p(\Delta r) = p(\Delta X)p(\Delta Y)p(\Delta Z) = \frac{1}{\sqrt{(2\pi)^{3N}|\frac{k_B T}{\gamma}\Gamma^{-1}|^3}} exp\{-\frac{3}{2}(\Delta X^T(\frac{k_B T}{\gamma}\Gamma^{-1})^{-1}\Delta X)\} \tag{4.30}$$

Where the term $\frac{1}{\sqrt{(2\pi)^{3N}|\frac{k_BT}{\gamma}\Gamma^{-1}|^3}}$ is the normalisation constant and $\frac{k_BT}{\gamma}\Gamma^{-1}$ is the co-variance matrix. Mean square fluctuations (the expected values of residue fluctuations) and cross correlations between nodes, and thus, in the present case, between residues, are given by the diagonal and off-diagonal terms of the covariance matrix respectively:

$$< \Delta r_i^2 >= \frac{3k_BT}{\gamma}(\Gamma^{-1})_{ii} \tag{4.31}$$

$$< \Delta r_i \cdot \Delta r_j >= \frac{3k_BT}{\gamma}(\Gamma^{-1})_{ij} \tag{4.32}$$

Normal modes are obtained through diagonalisation of the Kirchoff matrix. Mode frequencies and mode shapes are given by eigenvalues and eigenvectors respectively.

The anisotropic network model differs from the Gaussian elastic network model in that it accounts for the directionality of fluctuations. The harmonic potential between two Cαs is given by the equation:

$$V_{ij}(r) = \frac{\gamma}{2}(|r_i - r_j| - |r_i^0 - r_j^0|)^2 \tag{4.33}$$

Where r is the ensemble of all positions. This can be summed over all pairs ij where $i \neq j$:

$$V(r) = \sum_{i=1}^{N}\sum_{j\neq i}^{N} V_{ij}(r) \tag{4.34}$$

The potential for small Cα displacements can be calculated using a Taylor expansion about r^0:

$$V(r) = V(r^0) + (r - r^0)^T D + \frac{1}{2}(r - r^0)^T H(r - r^0) \tag{4.35}$$

First and second order derivatives of the potential V are denoted by D and H respectively. As the reference structure is assumed to represent an energy minimum, $V(r^0)$ and D evaluate to 0, and

the equation reduces to:

$$V(r) = \frac{1}{2}(r - r^0)^T H (r - r^0) \tag{4.36}$$

Each element $H_{i,j}$ of the hessian matrix H, is a 3 by 3 matrix which holds anisotropic information on the orientation of the modes.

$$H_{ij} = \begin{cases} -\frac{\gamma_{ij}}{|r_i - r_0|^2}(r_i^0 - r_j^0)(r_i^0 - r_j^0)^T, & \text{if } i \neq j \ \& \ r_{ij} \leq r_c \\ 0, & \text{if } i \neq j \ \& \ r_{ij} > r_c \\ \sum_{j,j\neq i}^{N} H_{ij}, & \text{if } i = j \end{cases} \tag{4.37}$$

Similar to the Kirchoff matrix of the GNM model, the Hessian matrix can be diagonalised to reveal the eigenvectors or normal modes of the ANM. In comparison to GNMs, here each mode is described by a 3 component vector, which gives information on the directionality of the mode (Lezon et al., 2010).

4.2.8 Analyses

Features derived from elastic network models (ENMs) were calculated using the ProDy software, and are identical to those calculated by Ponzoni and Bahar (2018). It must be noted that we consider the Q(SASA) an ENM feature for the purposes of our work, to maintain consistency with the grouping of features used by the Bahar laboratory (Ponzoni and Bahar, 2018). The ENM-based features are as follows:

- Mean square fluctuations (MSF): square displacements at each position calculated from all modes of a Gaussian network model.
- Mechanical bridging score (MBS): this is a measure of the importance of each residue in maintaining a graph representation of the structure.
- Effectiveness(EFF)/sensitivity(SNS): parameters calculated from a perturbation response scanning matrix. The column average indicates the effectiveness of a residue in transmitting

allosteric signals to the rest of the protein. The row average indicates the sensitivity of a residue to deformations at other sites.

- Mechanical stiffness (STF): resistance of each residue pair to uniaxial tension. Averaged over all pairs involving a given residue.
- Quotient solvent accessible surface area [Q(SASA)]: the quotient of the solvent accessible surface area and surface area of the isolated residue. Calculated for the mutated WT position using the POPS software (Cavallo et al., 2003).

Similarly, a number of features were calculated from atomistic molecular dynamics trajectories using the software MDAnalysis (Michaud-Agrawal et al., 2011). 101 snapshots from each trajectory have been used for more computationally intensive analyses, whereas 5001 snapshots have been used where this is not an issue. Appendix A9 shows that 101 snapshots are sufficient to capture trends in those properties which are more computationally intensive to calculate. The analyses are as follows:

- Root mean square deviation (RMSD): root mean square deviation of Cα atoms from their position in the static wild-type (WT) structure, calculated for 5001 snapshots of the trajectory.
- Root mean square fluctuations (RMSF): root mean square fluctuations of Cα atoms, in comparison to their mean positions, calculated over 5001 snapshots of the trajectory.
- Number of hydrogen bonds (HB): the total number of hydrogen bonds calculated for 101 snapshots of the trajectory.
- Number of hydrogen bonds at the mutated position (HBpos): the number of hydrogen bonds involving the amino acid at the position of the mutant, calculated for 101 snapshots of the trajectory.
- Number of contacts at the mutated position (CT): number of Cβs within 8 Å of the residue at the mutated position (Cαs are used to represent glycines).

Two matrix properties of the trajectories were also calculated using the software Bio3D (Grant et al., 2006) and GSATools respectively (Pandini et al., 2013):

- Contact map (CM): contact map with each value giving the frequency a particular contact is observed throughout the trajectory. Contacts are defined based on Cβs and a cut-off value of 8 Å. Calculated for 5001 snapshots of the trajectory.

- Mutual information (MI): For each trajectory snapshot, 4 residue fragments are transformed into a structural alphabet letter, based on the relative orientations of their alpha carbons. This is calculated for $n - 3$ overlapping fragments, where n is the number of residues the protein consists of. Thus each snapshot is represented as a string, and these strings form an alignment from which the normalised mutual information is calculated.

Additionally, dynamic graph-based features of the trajectories were calculated using the package MD-TASK (Brown et al., 2017). Here each snapshot of the trajectory is transformed into graph representation, in which the Cβs are represented by nodes, and those within 8 Å of one another are connected by edges. The following features are then calculated from these graphs (specifically, both the mean and standard deviation of these features are calculated per residue):

- Betweenness centrality (BC): the protein structure is transformed into a graph, where each residue is represented by a node positioned at its Cβ atom (Cα for glycines), and the betweenness centrality of each residue is calculated for 101 snapshots from the trajectory.

- Change in betweenness centrality (delta_BC): as for betweenness centrality but calculated as the difference in betweenness centrality between the initial trajectory frame in comparison to each of the 100 non-initial trajectory snapshots.

- Average shortest path (L): as for betweenness centrality, however here the average shortest path from one residue to all other residues is calculated.

- Change in the average shortest path (delta_L): as for the average shortest path, however here the difference between the initial trajectory frame in comparison to each of the 100 non-initial trajectory snapshots is calculated.

All dynamics properties calculated per residue (e.g. RMSF), or per frame (e.g RMSD) were transformed into numeric features, based on minimal differences between the mutant and two

wild-type replicas. First d_i, which quantifies this difference at frame/position i is extracted for each residue/frame (see Fig. 4.4):

$$d_i = \begin{cases} WT_{1_i} - mut_i, & \text{if } |WT_{1_i} - mut_i| < |WT_{2_i} - mut_i| \ \& \ mut_i \neq median(WT_{1_i}, WT_{2_i}, mut_i) \\ WT_{2_i} - mut_i, & \text{if } |WT_{2_i} - mut_i| < |WT_{1_i} - mut_i| \ \& \ mut_i \neq median(WT_{1_i}, WT_{2_i}, mut_i) \\ 0, & \text{if } mut_i = median(WT_{1_i}, WT_{2_i}, mut_i) \end{cases} \tag{4.38}$$

Where WT_{1_i} and WT_{2_i} are values calculated from each wild-type replica for frame/residue i, and mut_i is the analogous value calculated from the mutant trajectory.

The two features F and F_{abs} are then calculated:

$$F = \sum_{i=1}^{n} d_i \tag{4.39}$$

$$F_{abs} = \sum_{i=1}^{n} |d_i| \tag{4.40}$$

Where, depending on the property, n is either the number of frames or the number of residues.

For matrix-based properties (MI and CM) the Spearman correlations between WT and mutant trajectories were used as features. Here the most positive pairwise correlation between a mutant and WT replicas is selected.

Two structural features were also calculated. These are:

- Betweenness centrality: as for dynamic features but calculated only for the mutated position of the WT at the initial time point after equilibration. As two replicas are available the average of both replicas is used.

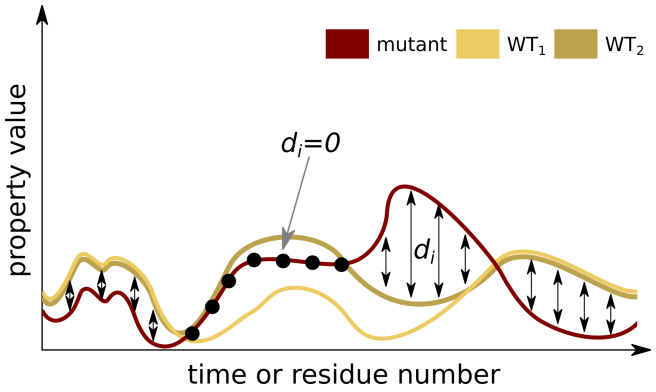

Fig. 4.4 Extraction of dynamics-based features. Properties are calculated over multiple time frames or residues, and the minimum differences (d) between wild-type and mutant trajectories are extracted.

- Average shortest path: as for betweenness centrality, however here the average shortest path from the WT mutated residue to all other residues is calculated.

As well as structure-based features, sequence-based features are also calculated as follows:

- Position Specific Independent Counts score (PSIC): This indicates the likelihood of observing an amino acid at a specific position and compares it to the background probability of observing it at any position (Sunyaev et al., 1999).
- Kidera factors (KF): 10 orthogonal properties which describe the physicochemical attributes of amino acids (Kidera et al., 1985).

Here both the WT properties, and differences between the WT and mutant properties, were calculated.

In addition to the calculation of features, principal component analysis was also performed using 101 snapshots of each trajectory, reduced to a Cα representation, using the ProDy software (Ponzoni and Bahar, 2018).

4.2.9 Random forest classifier

A random forest-based model, implemented in Scikit learn (Pedregosa et al., 2011), was trained to classify SAVs as deleterious or neutral. Random forests were chosen due to their relative robustness

and ability to deal with high dimensional data (small numbers of samples and large numbers of features) (Breiman, 2001). Due to the small size of the dataset Leave One Out (LOO) cross-validation was used to evaluate the performance of the classifier. Where > 6 features were available to build a model, a model was first created using all features and trained on the training fold. The top six most important features from this model were used to create a new model, again trained on the training fold. The performance of this model was then evaluated based on its predictive performance on the testing fold. Additionally, models were created which used the top six sequence-based features and the top six dynamics-based features (selected based on two separate models trained on the training fold). Again these models were trained on the training fold and evaluated based on their performance on the testing fold. The random forests were created using an ensemble of 1000 trees, and the maximum number of features to consider when splitting a node of a tree was set to 2. The "balanced" option of the SciKit learn implementation was chosen, which assigns class weights inversely proportional to their size in the training set. In this case, the option will have had minimal impact on the algorithm, as the benchmark dataset we use is close to being balanced.

4.2.10 Data visualisation

Data were visualised using the Python package matplotlib (Hunter, 2007). Structures were visualised using the VMD (Humphrey et al., 1996) and UCSF Chimera (Pettersen et al., 2004) software packages.

4.3 Results

4.3.1 Comparison between coarse-grained and atomistic simulations

The root mean square fluctuations of wild-type (WT) elastic network models, coarse-grained simulations and atomistic simulations positively correlate with one another, with values of Spearman's $\rho \geq 0.4$ (see Fig. 4.5). Moreover, atomistic molecular dynamics simulations show Spearman correlations between 0.7 and 0.93. The lowest correlation (0.7) between replicas is demonstrated by the domain Fn3-119. In-house experiments [2] show that this domain has a comparatively low melting temperature (50^oC). Furthermore, the fact that it has not been possible to crystallise this domain suggests that it is highly flexible. This flexibility is further supported by the RMSF values of our atomistic MD

simulations (see Fig. 4.14). Because of this greater flexibility, it is likely that a smaller portion of the conformational landscape accessible to this domain can be sampled during our 100 ns simulations. Due to this, it appears that each replica explores a different basin of the energy landscape, leading to smaller correlations between replicas.

Although properties of wild-type dynamics are captured through coarse-grained simulations, it is clear these techniques are not able to capture differences between mutations. This is likely due to the fact that the coarse-grained techniques used here model all forces in the "native" structure as attractive (via the harmonic restraints of ENMs and Martini ElNeDyn models and GO potentials of the CafeMol and Martini GO models). For computationally modelled mutant structures the assumption that the input structure represents an energy minimum is likely to break down. In comparison, clear differences can be seen between WT and mutant atomistic MD trajectories for a large proportion of mutants, as evidenced by RMSF values (see Figs 4.6, 4.7, 4.8, 4.14).

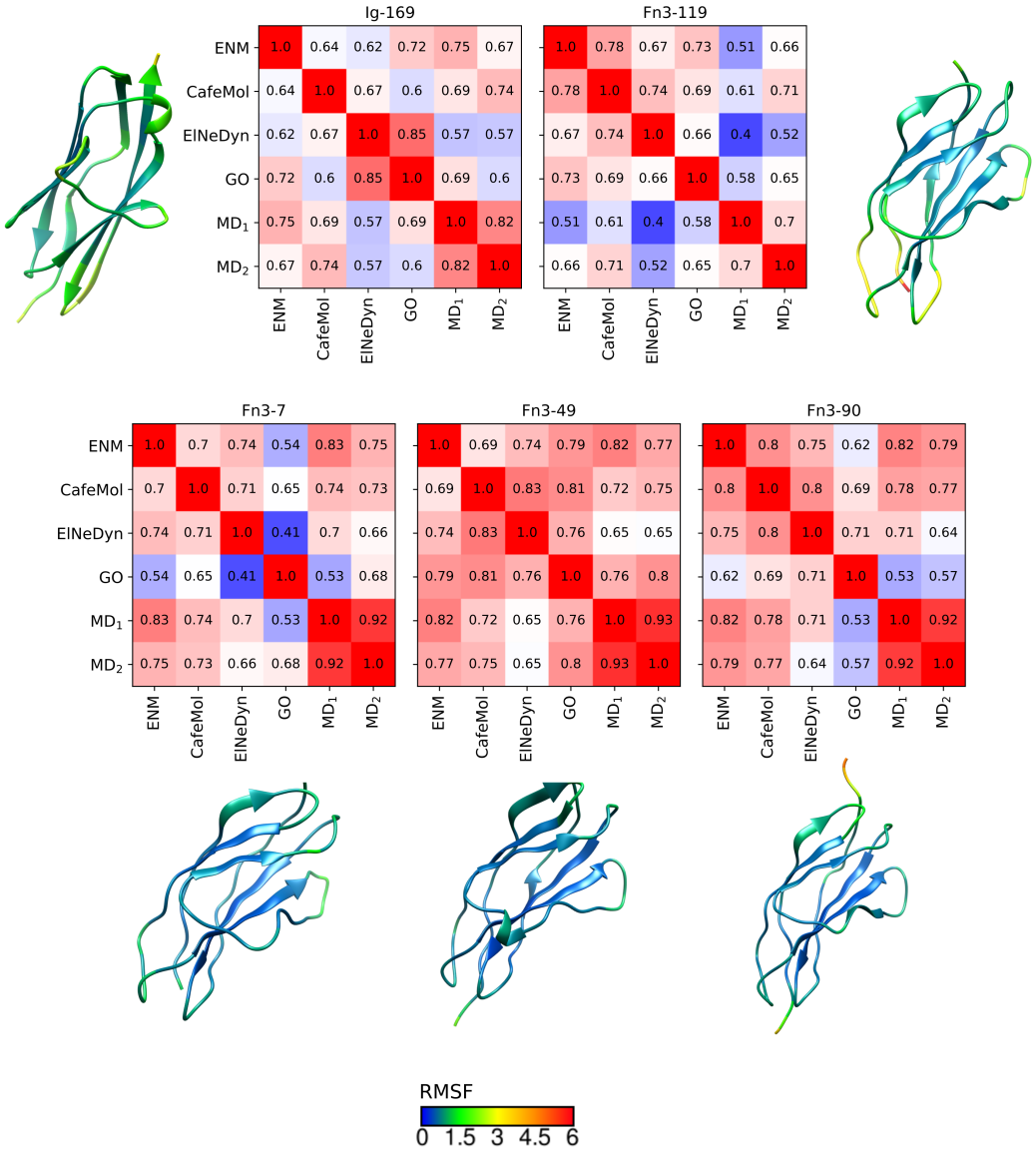

Fig. 4.5 Spearman correlations between RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics (MD) models. The heatmaps are both coloured and labelled by Spearman's ρ and show results for the wild-type (WT) titin domains Ig-169, Fn3-119, Fn3-7, Fn3-49 and Fn3-90. Two replicas are shown for atomistic molecular dynamics simulations (MD_1 and MD_2). Structures are coloured by MD_1 RMSF values and visualised using the UCSF Chimera software (Pettersen et al., 2004). Positive correlations can be seen for all models, showing that coarse-grained models are, to varying degrees, able to reflect WT dynamics. However, the largest correlations are always seen between atomistic MD replicas.

4.3.2 Variant analysis - atomistic simulations

For the domains Ig-169, Fn3-90, and Fn3-49, disease-associated mutant trajectories show higher RMSF values than population mutant trajectories. The single exception to this is the buried Ig-169 population variant (I55T), which appears to have a drastic impact on the domain's dynamics, leading to increased flexibility (see Fig. 4.6) and, what appears to be, the initialisation of unfolding. Interestingly this mutant is in the same position as the I55N TMD associated "Belgian" mutant. It is also extremely rare (gnomAD MAF 4.08E-06). Therefore we hypothesise this mutant may also lead to a disease-associated phenotype in a homozygous or compound heterozygous configuration. Interestingly whether the "Belgian" mutant is causative of the disease TMD is a matter of debate. Although cosegregation with the disease phenotype is observed, biophysical studies uncovered only small differences between the wild-type and mutant (Rudloff et al., 2015). However, our results suggest that the protein dynamics are perturbed in a similar manner to other TMD-associated mutants. Replicas for the I55T mutant and the W38R disease-associated mutant support clear disturbance of the domain's dynamics.

The RMSF values for the dynamics simulations for the domain Fn3-90 mutants reflect in-house experimental data [2]. Here it can be seen that the disease-associated G88E mutant displays an altered RMSF profile, whereas the I14V population mutant shows dynamics similar to the wild-type. The G88E population mutant displays characteristics between these two extremes. Differential scanning fluorimetry (DSF) data [2] shows that the I14V mutant has a melting temperature of 59^oC, similar to the WT (60^oC). In contrast, the R78Q mutant has a lower melting temperature of 47^oC and the G88V mutant is unfolded.

Although the RMSF profiles for the WT and mutant domain Fn3-49 do not, at first glance, appear markedly different, a clear distinction can be seen at the C-terminal end, between residues 90 and 100 (see Fig. 4.8). Here the disease-associated mutant shows increased flexibility in comparison to both WT replicas and the population mutant.

Principal component analyses also show clear distinctions between population and disease-associated mutant trajectories for these three domains (Ig-169, Fn3-90, and Fn3-49) (see Figs 4.9-4.11). Here WT and population mutant trajectories appear confined to distinct regions of the plots, whereas

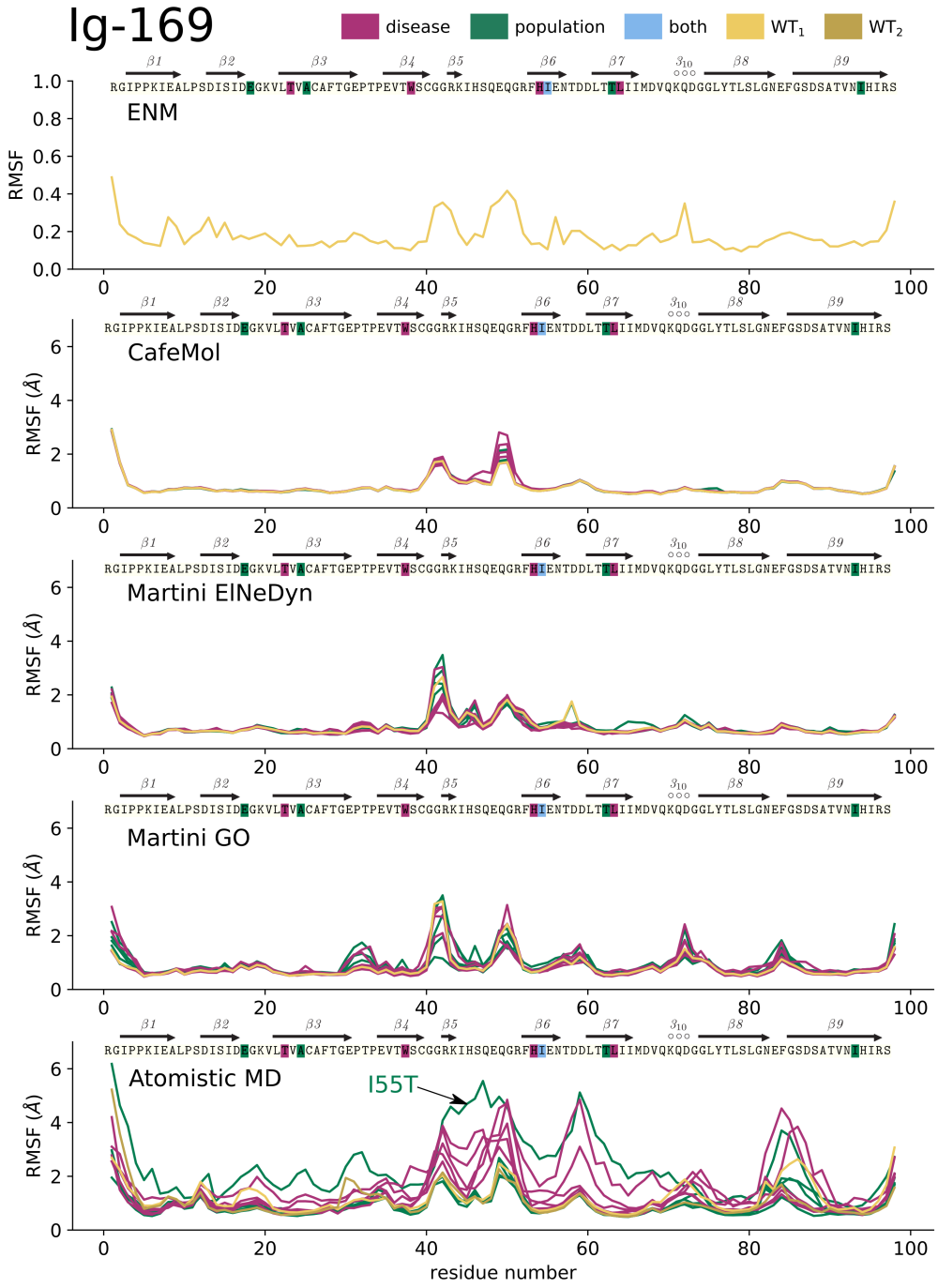

Fig. 4.6 Root mean square fluctuations (RMSF) for the wild-type and mutant domain Ig-169. Both population (green) and disease-associated (magenta) mutants are depicted. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. The position 55, to which both a population-associated and a disease-associated mutation localise, is coloured blue on the amino acid sequence; however, these mutants are coloured according to their class (population/disease) on the RMSF plots. Note that the units for RMSF values from the Gaussian elastic network model (ENM) are relative/arbitrary (Bakan et al., 2011). It can be seen that only atomistic molecular dynamics simulations allow for a clear separation between population mutant and disease-associated mutant RMSF values, with the exception of the I55T population mutant, which has been labelled on the plot. Like the disease mutants, this mutant demonstrates higher RMSF values in comparison to the wild-type and other population mutants.

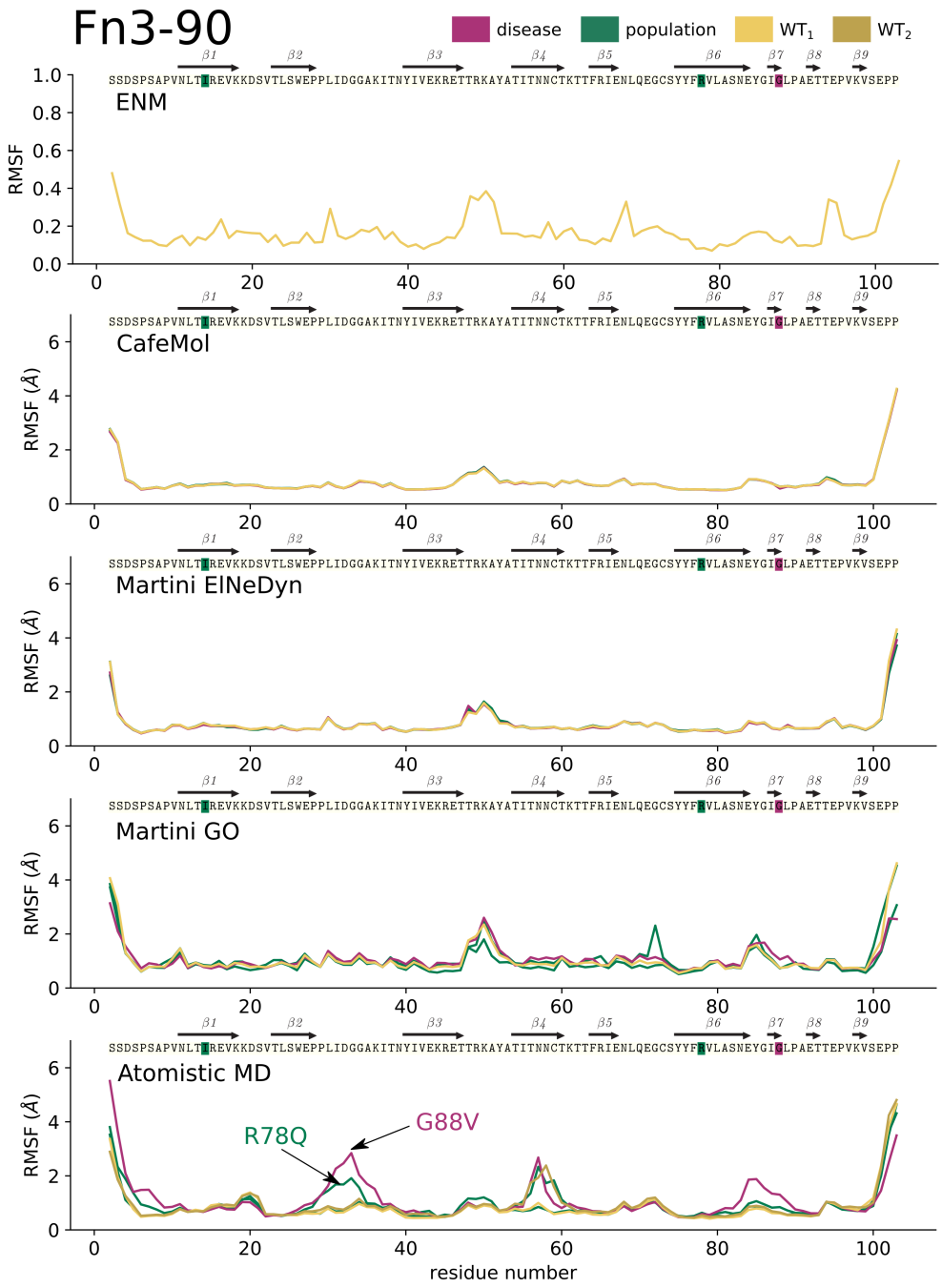

Fig. 4.7 Root mean square fluctuations (RMSF) for the wild-type and mutant domain Fn3-90. Both population (green) and disease-associated (magenta) mutants are depicted. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. Note that the units for RMSF values from the Gaussian elastic network model (ENM) are relative/arbitrary (Bakan et al., 2011). It can be seen that only atomistic molecular dynamics simulations allow for a clear separation between disease-associated and population mutant RMSF values. The G88V disease-associated mutant (labelled) shows the greatest difference in RMSF values from the wild-type. Interestingly the R78Q population mutant (labelled), which has been shown to be destabilised in comparison to the wild-type (WT) *in vitro* also exhibits greater RMSF values than the wild-type.

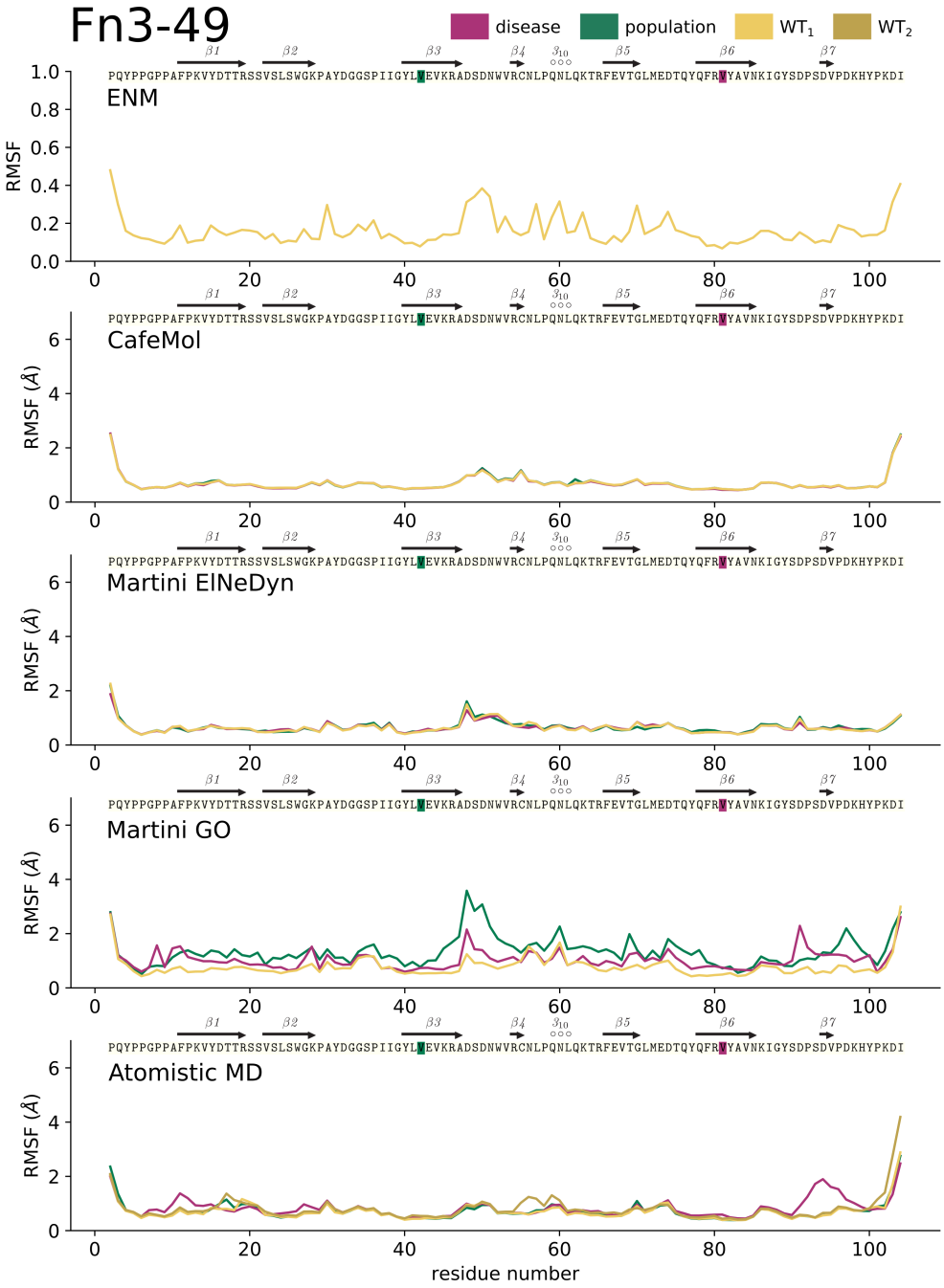

Fig. 4.8 Root mean square fluctuations (RMSF) for the wild-type and mutant domain Fn3-49. Both population (green) and disease-associated (magenta) mutants are depicted. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. Note that the units for RMSF values from the Gaussian elastic network model (ENM) are relative/arbitrary (Bakan et al., 2011). It can be seen that atomistic molecular dynamics simulations allow for separation of the disease-associated and population mutant trajectories, as the disease-associated mutant shows increased RMSF values between residues 90 and 100. The other models fail to achieve this separation.

the disease-associated mutants explore greater conformational space. Again the exception to this is the Ig-169 population I55T variant, for which both replicas follow a pattern closer to that of the disease-associated variants.

In contrast, the wild-type and mutant trajectories for the Fn3-7 domain show similar dynamic properties to one another (see Fig. 4.12). Moreover, principal component analysis shows a clear overlap between all mutant and WT trajectories projected onto the first two principal components (see Fig. 4.13). Interestingly this domain has a WT melting temperature 75^oC, which is much higher than the other studied domains (see Table 4.1); thus it is possible that a longer simulation time, or simulation at increased temperatures, would be necessary to observe perturbation of this domain by the disease-associated variant.

For the domain Fn3-119, variation in the RMSF profiles is seen between all simulations (see Fig. 4.14), however no segregation between population and disease-associated mutants is evident. Indeed, as has been discussed, even the WT replicas show higher variability than those of other domains. One plausible explanation is that the lack of segregation may be due to under-sampling of the conformational space available to this more flexible domain.

Principal component analysis for the domain Fn3-119 (see Fig. 4.15) shows an overlap in the trajectories of population variants and disease-associated variants when projected onto the first two principal components. Both WT trajectories appear confined to the top right-hand side of the plot. The only mutant with a similar localisation is the S59F mutant. Interestingly this is the variant with the highest MAF of those analysed here, and was found to have a thermal stability similar to the wild-type by the Gautel laboratory [2] (a melting temperature of 49^oC in comparison to 50^oC). The disease-associated C5R and P25L mutant trajectories localise to areas of the plot which diverge from the wild-type trajectories, and have been shown to have decreased thermal stability (32^oC and 33^oC respectively). These observations give rise to the hypothesis that a number of rare population-variants may have an impact on the protein dynamics and, therefore, function. Rather than the binary outcome suggested by the results for Ig-169, a spectrum of impacts may exist.

Another explanation for the lack of clear segregation between disease-associated and population mutant trajectories for the domains Fn3-7 and Fn3-119 could be attributed to the use of homology models rather than crystal structures. It is possible that any inaccuracies in these could impact on

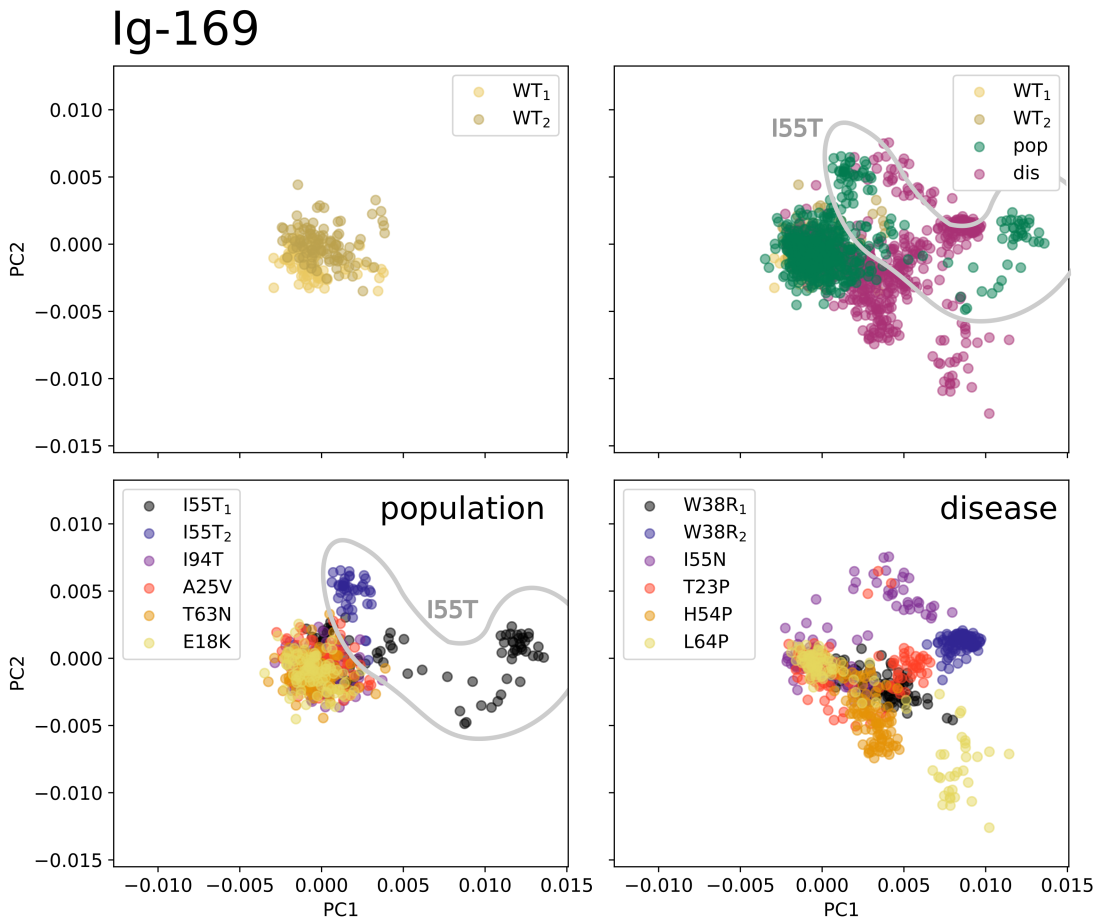

Fig. 4.9 PCA of atomistic MD trajectories for the wild-type and mutant Ig-169 domain. 101 snapshots of Cα coordinates from each trajectory have been taken for analysis. PC1 and PC2 explain 31 % and 9 % of the variance respectively. The top left plot shows the two wild-type replicas; the top right plot shows wild-type, disease-associated mutant and population mutant trajectories; the bottom left plot shows only population mutants, including two replicas for the I55T mutant; the bottom right plot shows only disease-associated mutants. The PCA analysis shows that wild-type and population mutant trajectories appear to be confined to a distinct region of the plot, whereas the disease-associated mutants appear to explore greater conformational space. The exception to this is the I55T mutant (outlined and labelled), for which both replicas follow a pattern more similar to disease-associated variants.

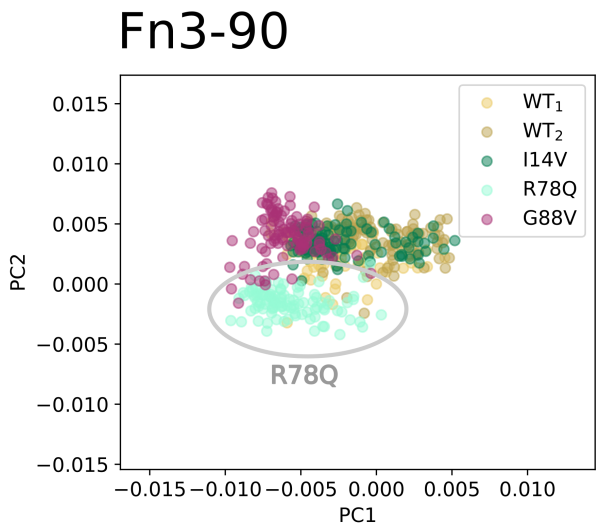

Fig. 4.10 PCA of atomistic MD trajectories for the wild-type and mutant Fn3-90 domain. 101 snapshots of Cα coordinates from each trajectory have been taken for analysis. PC1 and PC2 explain 24 % and 13 % of the variance respectively. Projections for two wild-type replicas, a disease-associated mutant (G88V) and two population mutants (I14V and R78Q) are shown. The I14V mutant appears confined to the region of the plot occupied by the two wild-type replicas, whereas the R78Q population mutant and G88V disease-associated mutant mainly occupy different areas of the plot. Interestingly the R78Q population mutant (outlined and labelled) has been shown to be destabilised in comparison to the wild-type (WT) *in vitro*.

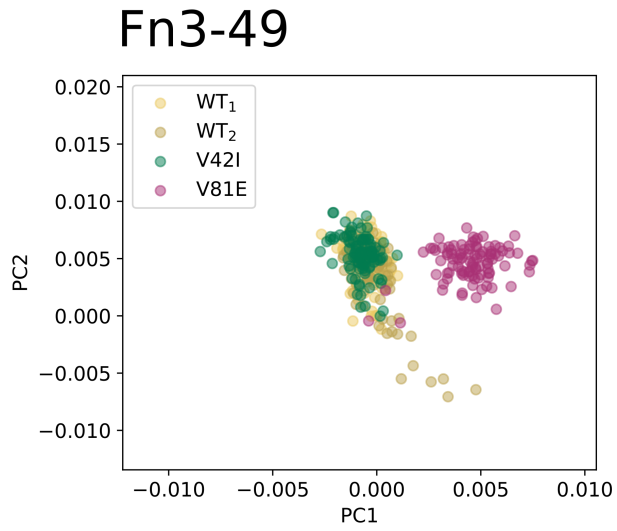

Fig. 4.11 PCA of atomistic MD trajectories for the wild-type and mutant Fn3-49 domain. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. 101 snapshots of Cα coordinates from each trajectory have been taken for analysis. PC1 and PC2 explain 20 % and 15 % of the variance respectively. Projections for two wild-type replicas, a disease-associated mutant (V81E) and a population mutant (V42I) are shown. The population mutant and wild-type replicas overlap and occupy a similar region of the plot whereas the disease-associated mutant shows little overlap with these.

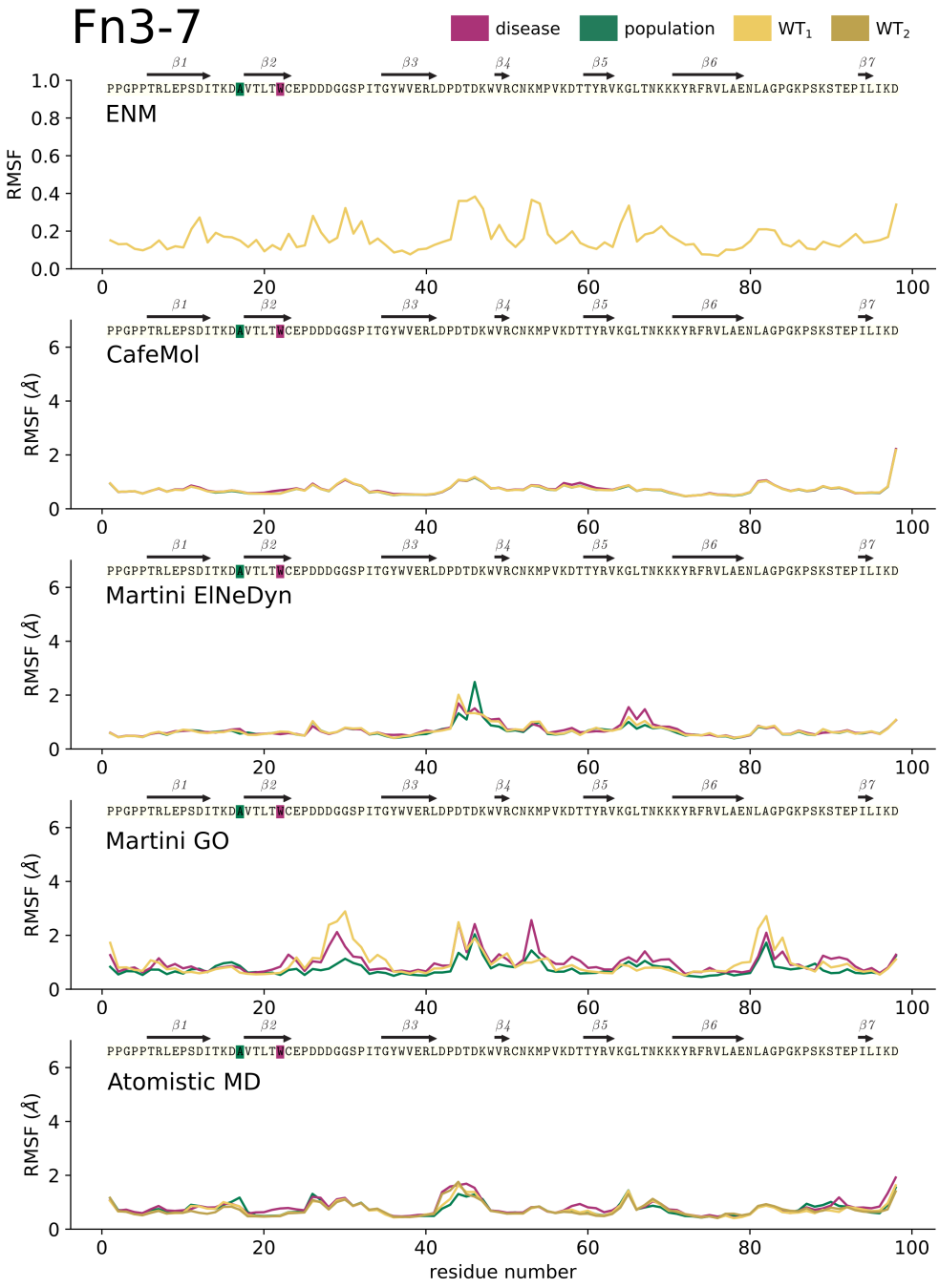

Fig. 4.12 Root mean square fluctuations (RMSF) for the wild-type and mutant domain Fn3-7. Both population (green) and disease-associated (magenta) mutants are depicted. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. Note that the units for RMSF values from the Gaussian elastic network model (ENM) are relative/arbitrary (Bakan et al., 2011). No clear differences between RMSF values for the population mutant and disease-associated mutant can be seen.

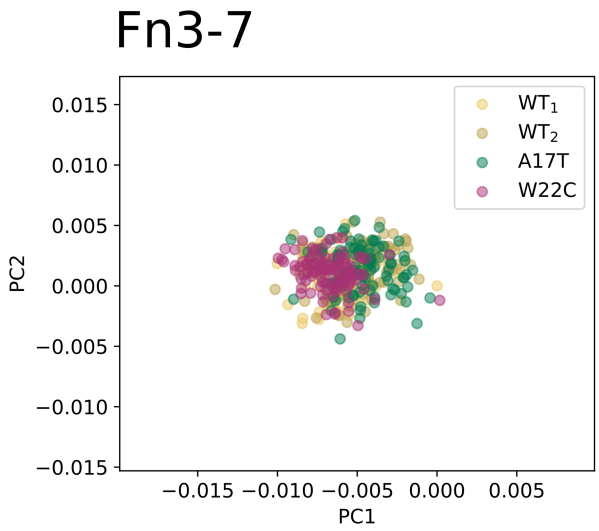

Fig. 4.13 PCA of atomistic MD trajectories for the wild-type and mutant Fn3-7 domain. 101 snapshots of Cα coordinates from each trajectory have been taken for analysis. PC1 and PC2 explain 13 % and 9 % of the variance respectively. Projections for two wild-type replicas, a disease-associated mutant (W22C) and a population mutant (A17T) are shown. The population mutant, disease-associated mutant and wild-type replicas overlap and occupy a similar region of the plot.

the simulation of dynamics. However, atomistic simulations using homology models for the WT and mutant Fn3-90 domain show good agreement with those simulations which use the in-house crystal structure as a starting point (see Fig. 4.16). It is important to note that information from this crystal structure was not used in the creation of the homology model.

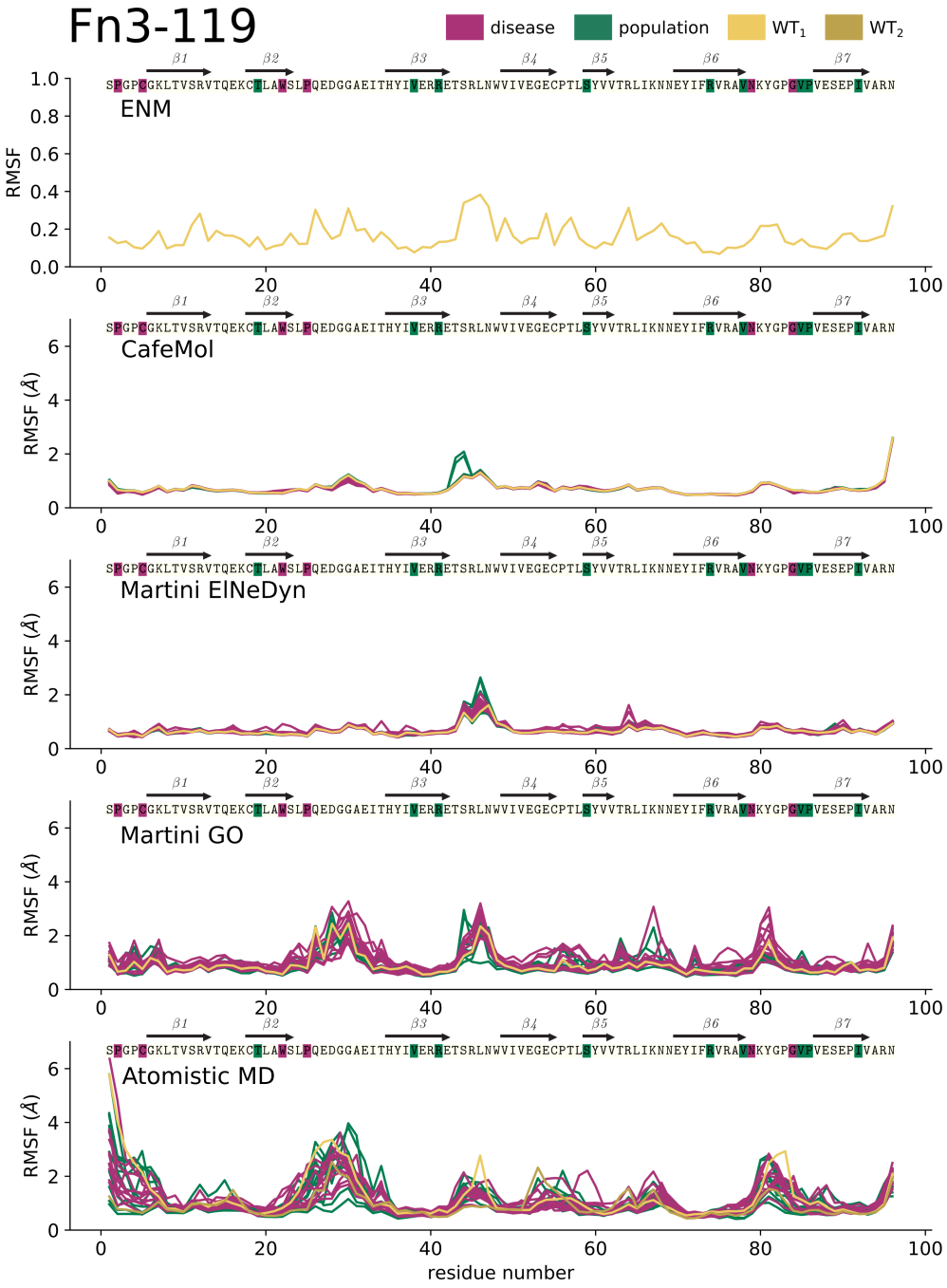

Fig. 4.14 Root mean square fluctuations (RMSF) for the wild-type and mutant domain Fn3-119. Both population (green) and disease-associated (magenta) mutants are depicted. Plots show RMSF values for elastic network, CafeMol, Martini ElNeDyn, Martini GO and atomistic molecular dynamics models. Note that the units for RMSF values from the Gaussian elastic network model (ENM) are relative/arbitrary (Bakan et al., 2011). Differences between the RMSF values for mutants are shown by both the Martini GO and atomistic molecular dynamics simulations, however no systematic distinction between disease-associated and population mutants is evident.

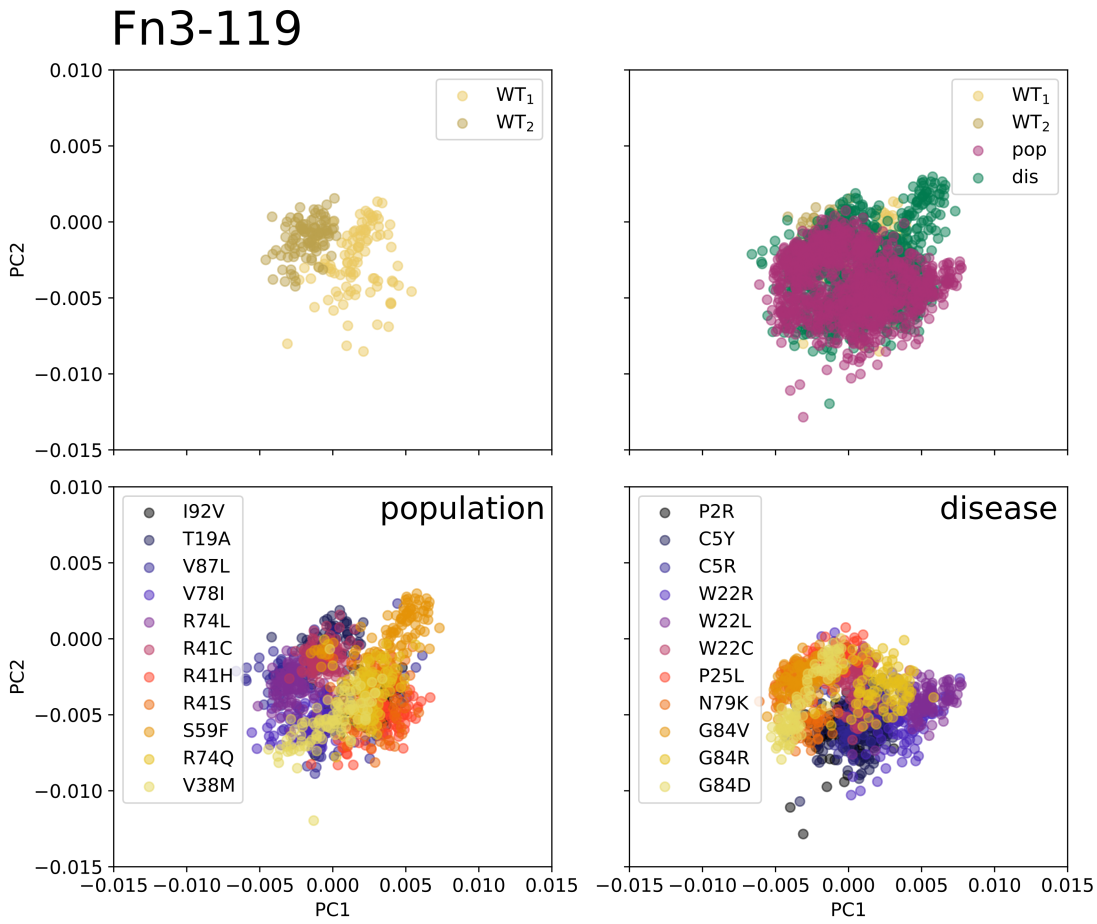

Fig. 4.15 PCA of atomistic MD trajectories for the wild-type and mutant Fn3-119 domain. 101 snapshots of Cα coordinates from each trajectory have been taken for analysis. PC1 and PC2 explain 15 % and 10 % of the variance respectively. The top left plot shows the two wild-type replicas; the top right plot shows wild-type, disease-associated mutant and population mutant trajectories; the bottom left plot shows only population mutants; the bottom right plot shows only disease-associated mutants. The PCA analysis shows that population mutant and disease-associated mutant trajectories overlap and diverge from the wild-type replicas, which appear confined to the top right-hand side of the plot.

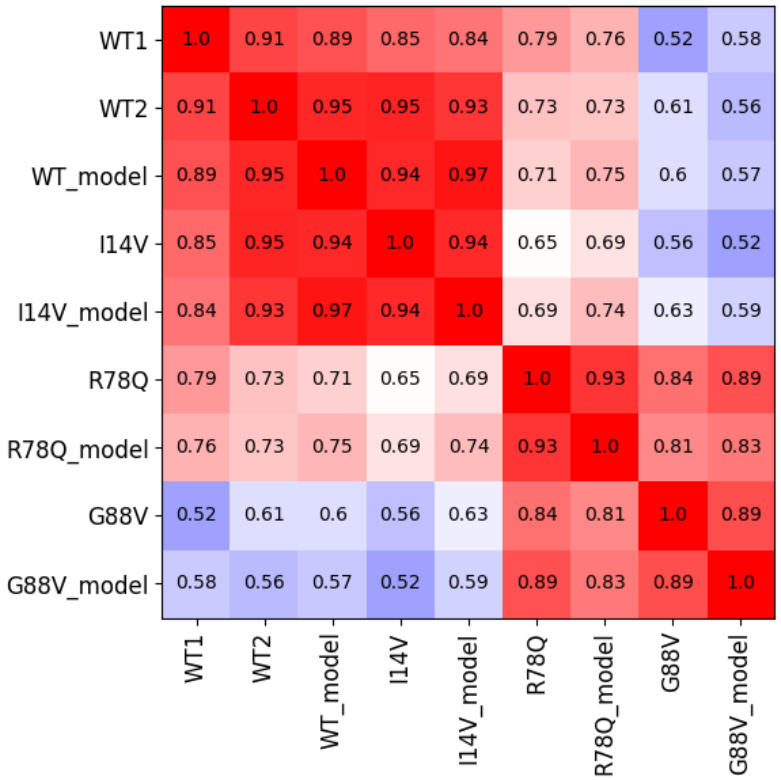

Fig. 4.16 Spearman correlations between RMSF values for mutant and wild-type (WT) atomistic molecular dynamics trajectories for the domain Fn3-90. The simulations have been performed using both a crystal structure and a homology model (_model) as a starting structure. The heatmap is both coloured and labelled by Spearman's ρ. Results are shown for both population mutants (I14V and R78Q) and a disease-associated mutant (G88V). Clear correlations are seen between WT replicas and between matched simulations (of the same mutant or WT) starting with coordinates from a model or a crystal structure. It can be seen that the I14V population mutant trajectories are highly correlated with all WT trajectories, whereas the G88V disease-associated mutant trajectories are the least correlated with these. The R78Q population mutant trajectories show correlations which are between those of the I14V population mutant trajectories and the G88V disease-associated mutant trajectories. Interestingly the R78Q population mutant has been shown to be destabilised in comparison to the wild-type *in vitro*.

4.3.3 Variant analysis - elastic network models

On comparison of features derived from elastic network models for the mutant subsets, it is clear there is a large overlap between the distributions for both the hotspot domains Ig-169 and Fn3-119 (see Fig. 4.17 and Fig. 4.18). Moreover, even if small differences can be detected between these distributions, it is clear that no robust segregation can be achieved using any of these calculated features alone. It is important to stress again here that these features only contain information about the dynamics of the WT position and not information pertaining to the physicochemical nature of the change.

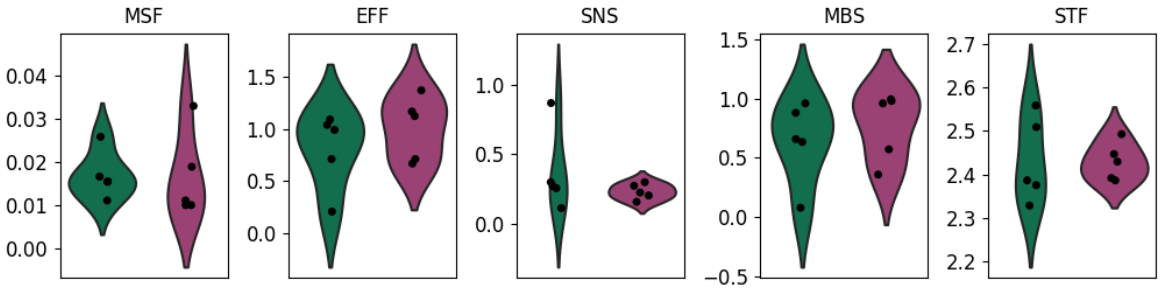

Fig. 4.17 Violin plots of ENM-based features calculated for population variants (green) and disease-associated variants (magenta) which localise to the domain Ig-169. Mean square fluctuations (MSF), effectiveness (EFF), sensitivity (SNS), mechanical bridging score (MBS) and mechanical stiffness (STF) are shown. All features are measured in arbitrary/relative units (Bakan et al., 2011). It can be clearly seen that the distributions of these features for population variants and disease-associated variants are largely overlapping.

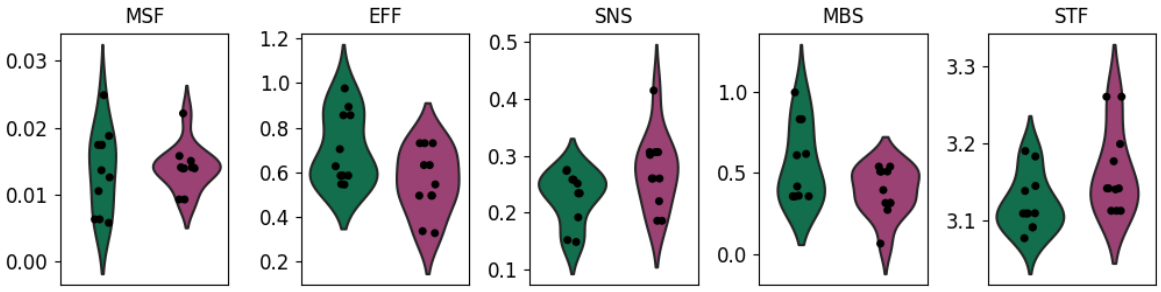

Fig. 4.18 Violin plots of ENM-based features calculated for population variants (green) and disease-associated variants (magenta) which localise to the domain Fn3-119. Mean square fluctuations (MSF), effectiveness (EFF), sensitivity (SNS), mechanical bridging score (MBS) and mechanical stiffness (STF) are shown. All features are measured in arbitrary/relative units (Bakan et al., 2011). It can be clearly seen that the distributions of these features for population variants and disease-associated variants are largely overlapping.

4.3.4 The predictive power of dynamics-based features

Here we compare the predictive power of features extracted from atomistic molecular dynamics simulations and ENMs in comparison to sequence-based features. To do this we create predictive random forest-based models, and assess their performance using Leave One Out (LOO) cross-validation. As described in the methods, the ENM-based features are identical to those calculated by Ponzoni and Bahar (2018). As 5 ENM-based features are calculated in their work, in addition to solvent accessible surface area, we similarly select the top 6 features according to their importance calculated using the training data, to create our atomistic dynamics- and sequence-based models (denoted RF_MD and RF_seq). We also create models allowing selection of the top 6 features from all calculated features (RF_all$_6$), including the two structure-based static network features, as well as models which use the top 6 sequence-based features and the top 6 dynamics-based features (RF_all$_{12}$). In addition to using the whole dataset to benchmark the models, we create models using the subset of the data for which crystal structures are available. This is to account for the possibility that simulations which use homology models as starting structures may be less accurate at reflecting true protein dynamics.

Based on the whole dataset, the atomistic dynamics- and sequence-based predictors are better able to segregate variants than the ENM-based predictors, according to all calculated metrics (see Table 4.3). Actually, the sequence-based predictor (F1 score 0.89) performs better than the dynamics-based predictor (F1 score 0.78), however both predictors show identical recall (0.84). From this analysis, it becomes clear that the perceived difference in performance is due to a greater number of false positives predicted by the dynamics-based model.

Using the subset of the data with crystal structures, both the sequence-based and atomistic dynamics-based models achieve even better segregation, whereas the ENM-based model performs badly (see Table 4.4 for performance metrics). The sequence-based model achieves perfect prediction; in comparison, the dynamics-based predictor performs almost as well and only misclassifies 2 population variants. Interestingly one of these variants is the Ig-169 I55T variant, which, as discussed earlier, we suspect could lead to a disease phenotype, due to both its rarity and localisation to the same position as the known "Belgian" disease-associated mutation. The other misclassified population

variant is the Fn3-90 R78Q mutant, which as discussed earlier has reduced thermal stability in comparison to the wild-type [2]. Although we cannot conclude that either variant is disease causing, the evidence suggests that these may not be functionally neutral. Therefore it is possible that the dynamics-based predictor could actually be describing a more accurate decision boundary with regards to whether variants have a functional impact.

In comparison to existing predictors, including PolyPhen2 (Adzhubei et al., 2010), Condel (González-Pérez and López-Bigas, 2011) and FATHMM (Shihab et al., 2013), both our sequence- and dynamics-based predictors show good performance, with the sequence-based predictor achieving the best performance across the majority of metrics (see Tables 4.3 and 4.4). The predictor REVEL also shows good performance on this dataset (Ioannidis et al., 2016). As this predictor has been specifically trained to separate rare neutral from disease-associated variants, it is perhaps unsurprising that it outperforms the majority of other predictors on this dataset. For a number of predictors a better ROC-AUC score is achieved than scores for other metrics. This indicates that the features calculated by these predictors have discriminative power, however the chosen cut-off for distinguishing disease-associated from neutral variants is not optimised for distinguishing rare neutral from pathogenic variants. It must be considered that the performance of these predictors may be inflated, as it is possible that they contain a subset of variants from our dataset in their training data. This is particularly likely in the case of REVEL, which was trained using rare variants. Moreover, as REVEL and Condel are meta-predictors, these could be subject to inflated performance due to the problem of type 1 circularity (see Section 1.4.1). However, we must be cautious when drawing conclusions about the relative performance of predictors; due to the small size of our benchmark dataset and the high variance of the LOO cross-validation procedure used to assess our predictor. If another set of 39 variants was used to benchmark the predictors, it is likely that their relative performance would change. Moreover, as our predictors have been trained using only titin variants in 5 titin domains, it is likely that our predictors are overfitting the data, i.e. they may not be able to generalise to variants in other proteins/domains.

Feature importances from predictive models that have been trained on the entire data, in addition to the number of times features are chosen during the LOO cross-validation procedure, are depicted in Fig. 4.19. This shows that, based on all structures, the Δ PSIC score has the most predictive

power, closely followed by dynamics-based features and the Kidera factors 1, 3 and Δ 4. The most important dynamics-based features appear to be associated with contacts (both those of the mutated position and global contact maps) and network-based metrics related to shortest paths. For the model based on the crystal structure data set, the Δ PSIC displays the highest importance, closely followed by dynamics-based features, with those associated with shortest path lengths playing the most prominent role. This suggests that the disruption of dynamic communication between residues, as measured by changes in shortest path lengths between WT and mutant trajectories, is a key determinant of variant pathogenicity.

Predictor	Accuracy	Precision	Recall	F1 score	MCC	ROC-AUC	Average precision
RF_seq	0.9	0.94	0.84	0.89	0.8	0.93	0.93
RF_all_{12}	0.85	0.84	0.84	0.84	0.69	0.91	0.91
REVEL (Ioannidis et al., 2016)	0.79	0.71	0.94	0.81	0.63	0.96	0.95
RF_all_{6}	0.79	0.82	0.74	0.78	0.59	0.81	0.83
RF_MD	0.77	0.73	0.84	0.78	0.55	0.81	0.71
MutationAssessor (Betts et al., 2015)	0.72	0.63	0.94	0.76	0.51	0.91	0.93
FATHMM_U (Shihab et al., 2013)	0.74	0.67	0.89	0.76	0.52	0.85	0.88
Condel (González-Pérez and López-Bigas, 2011)	0.77	0.74	0.78	0.76	0.54	0.92	0.93
Provean (Choi et al., 2012)	0.64	0.56	1	0.72	0.43	0.93	0.92
PPH2 (Adzhubei et al., 2010)	0.54	0.5	0.94	0.65	0.2	0.7	0.6
MutationTaster (Schwarz et al., 2010)	0.49	0.47	1	0.64	0.15	0.68	0.61
RF_ENM	0.59	0.58	0.58	0.58	0.18	0.65	0.73
FATHMM_W (Shihab et al., 2013)	0.64	1	0.22	0.36	0.37	0.74	0.79

Table 4.3 Performance of predictors on titin missense variants which localise to crystal structures and homology models (n=39). RF_seq, RF_all, RF_MD and RF_ENM are the predictors we have trained using sequence-based, mixed, atomistic dynamics and ENM-based features. The predictors Condel and REVEL are both meta-predictors.

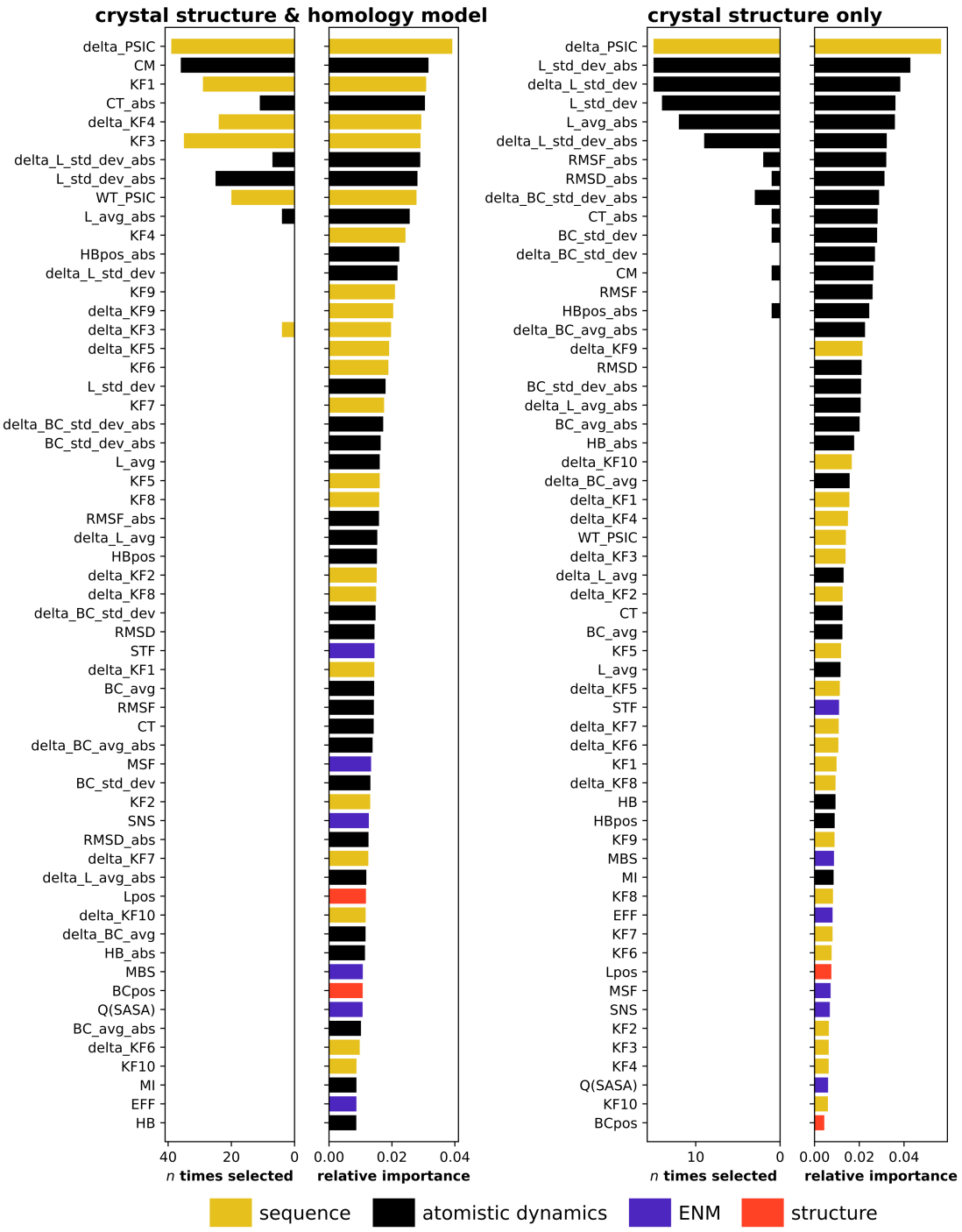

Fig. 4.19 The importance of different features for predicting the impact of SAVs using random forest-based models. The number of times a feature is selected during the cross-validation procedure is indicated on the left-hand side of each plot and feature importances are depicted for a model trained using all calculated features in the right-hand side of the plot. Shown for models trained on 39 variants (localised to crystal structures and homology models) and a subset of 15 variants (only crystal structures). Features have been described in Section 4.2.8

Predictor	Accuracy	Precision	Recall	F1 score	MCC	ROC-AUC	Average precision
RF_seq	1	1	1	1	1	1	1
REVEL (Ioannidis et al., 2016)	0.93	1	0.86	0.92	0.87	1	1
RF_all_6	0.87	0.78	1	0.88	0.76	0.93	0.92
RF_all_{12}	0.87	0.78	1	0.88	0.76	0.86	0.8
RF_MD	0.87	0.78	1	0.88	0.76	0.8	0.69
FATHMM_U (Shihab et al., 2013)	0.8	0.7	1	0.82	0.66	0.95	0.94
Provean (Choi et al., 2012)	0.73	0.64	1	0.78	0.56	1	1
Condel (González-Pérez and López-Bigas, 2011)	0.8	0.83	0.71	0.77	0.6	0.95	0.95
MutationAssessor (Betts et al., 2015)	0.73	0.67	0.86	0.75	0.49	0.93	0.94
PPH2 (Adzhubei et al., 2010)	0.6	0.54	1	0.7	0.37	0.88	0.88
MutationTaster(Schwarz et al., 2010)	0.53	0.5	1	0.67	0.25	0.65	0.61
FATHMM_W (Shihab et al., 2013)	0.6	1	0.14	0.25	0.29	0.77	0.81
RF_ENM	0.27	0.17	0.14	0.15	-0.49	0.16	0.35

Table 4.4 Performance of predictors on titin missense variants which localise to crystal structures only (n=15). RF_seq, RF_all, RF_MD and RF_ENM are the predictors we have trained using sequence-based, mixed, atomistic dynamics and ENM-based features. The predictors Condel and REVEL are both meta-predictors.

4.4 Discussion

Through this work, we have shown that features from molecular dynamics trajectories show promise for distinguishing between disease and population variants, and offer a clear performance advantage over elastic network derived models. Although our dynamics-based predictive models do not reach the performance of our sequence-based model, this is due to the misclassification of several population variants as disease-associated variants. Some evidence suggests that these "misclassified" variants may not be functionally neutral. This leads us to the hypothesis that dynamics-based features may be better at detecting functional impact in cases where sequence-based features fail. Moreover, we must consider that our disease-associated dataset may be biased towards variants which localise to conserved positions, as such variants are normally considered the most likely causative variants. Conversely, disease associations for variants which localise to less conserved positions may be more difficult to detect, and thus under-represented in our disease subset. Future experiments will be designed to probe whether this the case, and additionally whether our conclusions remain robust if these techniques are applied to larger datasets and different protein domain-types.

Coarse-grained techniques were able to produce reasonable representations of wild-type protein dynamics, but were unable to show differences between mutant trajectories. As already discussed, this is likely due to the restraints (elastic network and GO) which are required to maintain protein structure in the coarse-grained representation (WT Martini simulations without restraints resulted in a loss protein topology). It is possible that restraints could be relaxed or removed to a degree which would allow differences between wild-type and mutant domains to be seen, whilst still maintaining topology (unless a mutation is severely disruptive). If atomistic information is required, as our results suggest, the use of enhanced sampling techniques (such as Monte Carlo, replica exchange and simulated annealing (Hospital et al., 2015; van Gunsteren et al., 2018)) and/or implicit solvent (Kleinjung and Fraternali, 2014) could be investigated. These techniques could lower computational costs.

The degree to which rare variants have an impact on protein structure, and therefore function, is a matter of debate. Therefore, one of the aims of this work has been to gain a better understanding of this. We find the majority of rare variants show dynamical features which are highly similar to the wild-type protein, with only a minority showing properties which are clearly distinct. These variants we consider likely to have a functional impact, although whether such an impact is associated with disease cannot be discerned from the information available. Only for the domain Fn3-119 was large variety seen in the dynamic properties of rare variants, however clear differences were also seen between WT replicas. As we have already discussed, our results and in-house experimental data from the Gautel laboratory [2] suggest this domain is intrinsically more flexible. Therefore it may be necessary to conduct longer simulations to sufficiently sample conformational space. More generally, if molecular dynamics techniques are to be used on a large-scale for the assessment of variant impact, necessary simulation times must be benchmarked for different domain types. Here enhanced sampling techniques may play a role in sufficiently sampling conformational space (Hospital et al., 2015; van Gunsteren et al., 2018).

Simulations for the domain Fn3-90 using a crystal structure and a homology model show good agreement with one another. This suggests that simulations of high-quality homology models may accurately represent protein dynamics. However, the dynamics of the two domains for which only homology models are available (Fn3-7, Fn3-119) did not show clear segregation between

population and disease-associated variants. This could be due to inaccuracies in the homology models, under-sampling of conformational space, or an actual overlap in the impact of rare and disease-associated variants. An NMR structure for the domain Fn3-119 is currently being solved by the Gautel laboratory [2], and promises to shed light on this issue.

A number of disease-associated variants are known to impact on protein interactions. It would be interesting to take a sample of these and see how many can be correctly classified by studying the dynamics of a single domain/protein in isolation. It is likely that a fraction of variants prevent interactions, due to conformational changes, observable in the monomer, which impact on the binding site. Alternatively, some variants may have no impact on the conformation of the monomer, instead only exerting an impact on the binding site. The results from such an investigation could shed light on how broadly applicable the approach presented here is.

In conclusion, it is clear that protein dynamics can play an important role in understanding the impact of missense variants. The incorporation of the physicochemical nature of a mutation in dynamics simulations, and extraction of features associated with the differences between wild-type and mutant trajectories, appears to allow for a better distinction between pathogenic and non-pathogenic variants than WT dynamics features from Cα elastic network model representations. However, more work is needed to discern whether dynamics features may have an advantage over sequence-based features for classifying "problem case" variants. Finally, our results present scope for future development, benchmarking and application. We have shown dynamics features hold predictive power, now it is necessary to discern the best way to harness this.

Chapter 5

Discussion and perspectives

Despite technological progress, which enables the sequencing of human genomes at an unprecedented rate, we are still far from a full understanding of the genetic code. However, potential applications of such an understanding are manifold, and include drug development, precision medicine, diagnostics and prognostics. The limits of our current comprehension are highlighted by the problem of missing heritability, discussed throughout this work, where the genetic component of a phenotype remains unidentified. Variant impact predictors can play an important role in elucidating which variants lead to a given phenotype. Here, another limitation is encountered; a large proportion of current predictors rely on sequence-based evolutionary features, which give little insight into potential disease mechanisms. Therefore the overarching aim of this work was to gain a better understanding of which protein structural and functional properties could be of use in both the prediction and interpretation of the impact of genetic missense variants.

A specific purpose, here, was to contribute to improving the computational impact prediction of titin variants. These are of particular interest as they provide a prime example of variants which are difficult to classify using statistical methods, and thus necessitate the use of alternative means of assessment, such as computational impact predictors. We have facilitated the improved assessment of titin variants directly through the creation of TITINdb, a resource presented in Chapter 2, which allows users to access titin homology models, structures and missense variants. Additionally, this goal has been addressed in Chapter 4, through the creation of computational sequence- and dynamics-based predictors, for the assessment of titin missense variants. Furthermore, as these titin

missense variants localise to Fn3 and Ig domains, which are ubiquitous throughout the proteome, we believe this also addresses our overarching aim to uncover which properties can be used to predict the impact of missense variants. The large-scale analysis of missense variants in health and disease, presented in Chapter 3, further contributes to this goal. Although this analysis is not directly associated with titin variants, we believe the trends uncovered have implications for the assessment of all missense variants, including those which localise to the titin protein.

Active debate in the field is associated with the relative properties of rare and common variants. It has been both argued that common variants have more functional impact than rare variants, and that rare variants are more similar to disease-associated variants (Alhuzimi et al., 2018; Mahlich et al., 2017). Understanding this is of utmost importance to discerning which titin variants are likely to be disease-associated. Our research suggests that rare variants, overall, display properties between those of common and disease-associated variants, but are most similar to common variants. Zooming into the molecular scale, in particular on protein dynamics, our work suggests that the majority of rare titin variants have little functional impact, whereas a minority perturb dynamics to a similar degree as disease-associated variants. This supports the hypothesis that a small proportion of rare variants may act as phenotypic modifiers in particular constellations or possess undiagnosed disease associations.

The work we have presented in Chapter 3 highlights that greater understanding of the impact of SAVs can be gained by taking an integrative approach. Furthermore, recent developments in proteomics technologies can be harnessed to gain insight into the potential impact of variants in their cellular environment. Even since the completion of this work, new data has become available which probes changes in protein thermal stability throughout the cell cycle (Becher et al., 2018). From our analysis, it is clear that a complex interplay exists between the variant enrichment of proteins and their constituent structural regions, transcript expression, protein abundance, protein thermal stability and protein turnover. Future challenges will be associated with the further untangling of this data, to understand which correlations represent causal relationships. Although high throughput proteomics technologies are now able to take measurements for several thousand proteins, they have not yet obtained full coverage of the proteome. Attaining greater coverage would facilitate the use of such proteomics features in variant impact predictors.

Prediction and its assessment rely on accurate benchmark datasets. We show that predictors which have been trained on common variants do not choose an appropriate decision boundary for distinguishing rare neutral variants from rare pathogenic ones. Moreover, our incomplete understanding of which rare variants are functionally neutral, and biases in the detection of disease-associated variants, may impact on the accurate training and assessment of variant impact predictors. Emerging data in the field from saturation mutagenesis studies hold promise to aid in the clarification of these issues (Baugh et al., 2016; Findlay et al., 2018; Gray et al., 2018). This data will offer a gold standard with which to benchmark variant impact predictors in the future. However, it is important to note that saturation mutagenesis studies do not necessarily measure the impact of variants on all functions of a protein.

We demonstrate that atomistic protein dynamics show promise in the prediction of variant impact, and on our dataset of titin variants show clear performance advantages over dynamics-based features derived from elastic network models. Whether dynamics-based features offer advantages over sequence-based features remains to be seen. Although an initial assessment suggests that our sequence-based predictor outperforms that based on dynamics features, some evidence suggests that the misclassified false positives from our dynamics-based predictor may actually have a functional impact. As titin is, in essence, a polymer consisting of linked Fn3 and Ig domains, it is also possible that both predictions and insights into molecular disease mechanisms could be improved by simulating multidomain constructs. Moreover, the impact of variants on titin protein-protein interaction complexes should be considered. Unfortunately as described in Chapters 1 and 2, atomic details exist for only a few of titin's protein-protein interactions. However, elucidating these is an area of active research, so it is possible that more atomic-resolution data for such interactions will be available soon. As a starting point, the impact of Ig-169 variants on titin's interactions with obscurin and obscurin-like protein could be simulated. Despite titin's highly similar mode of interaction with both partners, subtle distinctions have been noted (Pernigo et al., 2015). Any differences observed between the impact of variants on interactions with each of these partners could shed light on disease mechanisms. As discussed in Chapter 4, this domain (Ig-169) is a hotspot for variants associated with the diseases TMD and LGMD-2J. Because we have simulated the impact

of the variants on the monomeric domain, it would be interesting to compare the results with simulations of variant impact on both complexes.

Our work suggests that sequence-based predictors perform well if trained on appropriate datasets. This raises the question, why should one invest in dynamics-based methods which have a much higher computational cost? There are several reasons for this. Firstly, we hypothesise that dynamic information may offer predictive power where sequence-based methods fail. Our results hint that this may be the case, for example the R78Q Fn3-90 titin variant (R27839Q based on IC isoform numbering) is predicted to be pathogenic by our dynamics-based predictor, but neutral by our sequence-based predictor. Although we cannot conclude this variant has any disease associations, experimental data show that this variant destabilises the Fn3-90 domain. We believe the next steps towards investigating this hypothesis should be to benchmark dynamic and sequence-based predictors on saturation mutagenesis datasets. In particular, it will be interesting to compare the performance of these predictors on rheostat positions, i.e. those protein positions for which a range of impacts exist (Miller et al., 2017). As discussed in the introduction, it has been shown that existing, primarily sequence-based, predictors perform poorly on variants which localise to these positions. Therefore these variants offer ideal test cases for dynamic methods. Secondly, we have to consider that this is a first attempt to use atomistic molecular dynamics-based features in a machine learning-based framework for the prediction of variant deleteriousness, whereas sequence-based methods are at a comparatively mature stage of development. Moreover, sequence-based methods incorporate alignments which utilise data from a much larger number of proteins. It is not clear how much dynamics-based methods could be improved given a larger training dataset and further development. Furthermore, it is also possible that dynamics-based information from homologs could further enhance performance. Thirdly, dynamic-based methods promise to give greater insight into molecular disease mechanisms than sequence-based methods. Additionally, simulation trajectories can be repurposed for use in drug screens and searches for potential ligand binding sites (Durrant and McCammon, 2011).

How is the use of atomistic molecular dynamics feasible for the assessment of variant impact on a large scale? Even with increased computational power and algorithmic improvements, which exploit parallel computing via the use of graphical processing units (GPUs) and potentially field-

programmable gate arrays (FPGAs) (Kutzner et al., 2015; Schaffner and Benini, 2018), proteome-wide assessment of SAVs using dynamics techniques is not feasible by any single research group. Rather, trajectories from multiple research groups could be pooled. We envisage the creation of an online database, akin to the protein data bank (PDB) to store and enable access to wild-type and mutant trajectories deposited by research groups across the world. Associated with this database, a machine learning-based variant impact predictor could iteratively learn from deposited data; this would enable the predictor to improve in parallel with the growth of the database. Therefore a group wishing to decipher the impact of a particular variant could simply run trajectories for the mutant protein and potentially a wild-type and neutral variant (in order to keep the database balanced) and submit these to the database to obtain the predictive results. Challenges here would lie in data curation, annotation and storage; however, these are not insurmountable given sufficient resources. Moreover, annotation of deposited trajectories with forcefield usage and simulation parameters would allow for additional benchmarking of the impact of these parameters on the predictive capacity of dynamics-based features.

Another consideration is that the proteome-wide assessment of dynamics is unlikely to be necessary. Those variants which cannot be classified based on functional, structural and proteomics features (as highlighted in Chapter 3) should be prioritised for dynamics investigation. This will increase the computational feasibility of the approach. Moreover, as discussed in Chapter 4 the use of enhanced sampling techniques (Hospital et al., 2015; van Gunsteren et al., 2018) and potentially implicit solvent (Kleinjung and Fraternali, 2014) may increase the efficiency of computational assessment. Additionally, a large database of mutant and wild-type trajectories might facilitate the design of coarse-grained methods which are able to capture the impact of mutations on protein dynamics.

To make full use of such a database, accurate metadata, both regarding patient phenotypes and available biophysical data describing WT and mutant proteins, would be necessary. The degree by which protein structure has to be disrupted to impact on function is not fully understood. Furthermore, the degree by which protein function must be perturbed to give rise to a disease phenotype is yet to be fully elucidated. However, to decipher how genetic variants lead to disease, we must understand both these links. The first steps towards such an understanding are initialised

by cataloguing available data. Unfortunately, despite the large amount of genetic data which can be accessed, the amount of individualised per-patient data, with associated phenotype information is much smaller. The safeguarding of patients, and associated controls regarding data sharing, whilst a necessity from an ethical perspective, can prove a hindrance to scientific advance (Raza and Hall, 2017). As an example as to how these problems can be overcome, the Danish healthcare system provides a paradigm for the notation, storage and access to patient data. Their detailed electronic records (notably opt-out rather than opt-in) have allowed researchers to create disease trajectories, which describe the temporal progression of diseases (Jensen et al., 2014). We believe that the association of such trajectories, with the impact of a patient's genetic variants at the molecular level, will be essential to gain a complete understanding of the link between phenotype and genotype. Furthermore, personalised phased genomic data is necessary if the problem of epistasis, the combinatorial impact of variants, is to be approached. Although databases such as gnomAD (Lek et al., 2016) represent a huge advance, these give us no information about which combinations of variants occur in an individual, and whether these occur in cis (on the same allele) or trans (on different alleles). This is of huge importance to understanding diseases with compound heterozygous inheritance.

In conclusion, through this work, we have shed light on a small part of the puzzle associated with deciphering the impact of genetic variants and, in the grand scheme of things, taken tiny but significant steps towards improving their assessment. Of greater importance are the questions this work leads us to ask and the future steps we believe need to be taken. Central to this is the sharing of data associated with the computational simulation of protein dynamics. It was exactly this open collaborative approach which enabled the advances of the genomic revolution and the reading of the genetic code (Contreras and Knoppers, 2018; Hood and Rowen, 2013). Now this approach is necessary in order for us to understand the genetic code, and ultimately devise our own instruction manual.

References

Acuna-Hidalgo, R., Veltman, J. A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome biology*, 17(1):241.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, Chapter 7:Unit7.20.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–9.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Banet, J. F., Billis, K., Girón, C. G., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., Vogel, J.-H., White, S., Zadissa, A., Flicek, P., and Searle, S. M. J. (2016). The Ensembl gene annotation system. *Database : the journal of biological databases and curation*, 2016.

Al-Numair, N. S. and Martin, A. C. R. (2013). The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC genomics*, 14 Suppl 3:S4.

Alexov, E. and Sternberg, M. (2013). Understanding molecular effects of naturally occurring genetic differences. *Journal of molecular biology*, 425(21):3911–3.

Alhuzimi, E., Leal, L. G., Sternberg, M. J. E., and David, A. (2018). Properties of human genes guided by their enrichment in rare and common variants. *Human mutation*, 39(3):365–370.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.

Ancien, F., Pucci, F., Godfroid, M., and Rooman, M. (2018). Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Scientific reports*, 8(1):4480.

Andersen, L. L., Terczyńska-Dyla, E., Mørk, N., Scavenius, C., Enghild, J. J., Höning, K., Hornung, V., Christiansen, M., Mogensen, T. H., and Hartmann, R. (2017). Frequently used bioinformatics tools overestimate the damaging effect of allelic variants. *Genes and immunity*.

Arimura, T., Bos, J. M., Sato, A., Kubo, T., Okamoto, H., Nishi, H., Harada, H., Koga, Y., Moulik, M., Doi, Y. L., Towbin, J. A., Ackerman, M. J., and Kimura, A. (2009). Cardiac ankyrin repeat protein gene (ANKRD1) mutations in hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, 54(4):334–42.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Azevedo, L., Mort, M., Costa, A. C., Silva, R. M., Quelhas, D., Amorim, A., and Cooper, D. N. (2016). Improving the in silico assessment of pathogenicity for compensated variants. *European journal of human genetics : EJHG*, 25(1):2–7.

Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology*, 21(3):432–40.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortés-Ciriano, I., Zhou, D. C., Liang, W.-W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Allen, E. M. V., Liang, H., Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R., and Ding, L. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18.

Bakan, A., Meireles, L. M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics (Oxford, England)*, 27(11):1575–7.

Bakir-Gungor, B., Egemen, E., and Sezerman, O. U. (2014). PANOGA: a web server for identification of snp-targeted pathways from genome-wide association study data. *Bioinformatics (Oxford, England)*, 30(9):1287–9.

Baresić, A., Hopcroft, L. E. M., Rogers, H. H., Hurst, J. M., and Martin, A. C. R. (2010). Compensated pathogenic deviations: analysis of structural effects. *Journal of molecular biology*, 396(1):19–30.

Baugh, E. H., Simmons-Edler, R., Müller, C. L., Alford, R. F., Volfovsky, N., Lash, A. E., and Bonneau, R. (2016). Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic acids research*, 44(6):2501–13.

Becher, I., Andrés-Pons, A., Romanov, N., Stein, F., Schramm, M., Baudin, F., Helm, D., Kurzawa, N., Mateus, A., Mackmull, M.-T., Typas, A., Müller, C. W., Bork, P., Beck, M., and Savitski, M. M. (2018). Pervasive protein thermal stability variation during the cell cycle. *Cell*, 173(6):1495–1507.e18.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS computational biology*, 10(1):e1003440.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature structural biology*, 10(12):980.

Betts, M. J., Lu, Q., Jiang, Y., Drusko, A., Wichmann, O., Utz, M., Valtierra-Gutiérrez, I. A., Schlesner, M., Jaeger, N., Jones, D. T., Pfister, S., Lichter, P., Eils, R., Siebert, R., Bork, P., Apic, G., Gavin, A.-C., and Russell, R. B. (2015). Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic acids research*, 43(2):e10.

Bienert, S., Waterhouse, A., de Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL repository-new features and functionality. *Nucleic acids research*, 45(D1):D313–D319.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, 40(6):695–701.

Bodmer, W. F. (1973). Genetic factors in hodgkin's disease: association with a disease-susceptibility locus (DSA) in the HL-A region. *National Cancer Institute monograph*, 36:127–34.

Bogomolovas, J., Gasch, A., Simkovic, F., Rigden, D. J., Labeit, S., and Mayans, O. (2014). Titin kinase is an inactive pseudokinase scaffold that supports MuRF1 recruitment to the sarcomeric M-line. *Open biology*, 4(5):140041.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brown, D. K., Penkler, D. L., Amamuddy, O. S., Ross, C., Atilgan, A. R., Atilgan, C., and Bishop, Ö. T. (2017). MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics (Oxford, England)*, 33(17):2768–2771.

Burley, S. K., Kurisu, G., Markley, J. L., Nakamura, H., Velankar, S., Berman, H. M., Sali, A., Schwede, T., and Trewhella, J. (2017). PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure (London, England : 1993)*, 25(9):1317–1318.

Butler, B. M., Gerek, Z. N., Kumar, S., and Ozkan, S. B. (2015). Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. *Proteins*, 83(3):428–35.

Campuzano, O., Sanchez-Molero, O., Mademont-Soler, I., Riuró, H., Allegue, C., Coll, M., Pérez-Serra, A., Mates, J., Picó, F., Iglesias, A., and Brugada, R. (2015). Rare titin (TTN) variants in diseases associated with sudden cardiac death. *International journal of molecular sciences*, 16(10):25773–87.

Canty, A. and Ripley, B. D. (2017). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-20.

Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P. L., Altman, R. B., and Casadio, R. (2013). WS-SNPsandGO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC genomics*, 14 Suppl 3:S6.

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, 33(Web Server issue):W306–10.

Carluccio, C., Fraternali, F., Salvatore, F., Fornili, A., and Zagari, A. (2013). Structural features of the regulatory ACT domain of phenylalanine hydroxylase. *PloS one*, 8(11):e79482.

Cavallo, L., Kleinjung, J., and Fraternali, F. (2003). POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic acids research*, 31(13):3364–6.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:7.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379.

Chauveau, C., Bonnemann, C. G., Julien, C., Kho, A. L., Marks, H., Talim, B., Maury, P., Arne-Bes, M. C., Uro-Coste, E., Alexandrovich, A., Vihola, A., Schafer, S., Kaufmann, B., Medne, L., Hübner, N., Foley, A. R., Santi, M., Udd, B., Topaloglu, H., Moore, S. A., Gotthardt, M., Samuels, M. E., Gautel, M., and Ferreiro, A. (2014a). Recessive TTN truncating mutations define novel forms of core myopathy with heart disease. *Human molecular genetics*, 23(4):980–91.

Chauveau, C., Rowell, J., and Ferreiro, A. (2014b). A rising titan: TTN review and mutation update. *Human mutation*, 35(9):1046–59.

Cheng, T. M. K., Lu, Y.-E., Vendruscolo, M., Lio', P., and Blundell, T. L. (2008). Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS computational biology*, 4(7):e1000135.

Chesmore, K. N., Bartlett, J., Cheng, C., and Williams, S. M. (2016). Complex patterns of association between pleiotropy and transcription factor evolution. *Genome biology and evolution*, 8(10):3159–3170.

Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science (New York, N.Y.)*, 265(5170):346–55.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688.

Chung, S. S., Laddach, A., Thomas, N. S. B., and Fraternali, F. (2018). Short loop motif profiling of protein interaction networks in acute myeloid leukaemia. *bioRxiv*.

Contreras, J. L. and Knoppers, B. M. (2018). The genomic commons. *Annual review of genomics and human genetics*, 19:429–453.

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., and Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. *Proteins*, 77 Suppl 9:18–28.

Das, S., Dawson, N. L., and Orengo, C. A. (2015). Diversity in protein domain superfamilies. *Current opinion in genetics and development*, 35:40–9.

David, A., Razali, R., Wass, M. N., and Sternberg, M. J. E. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human mutation*, 33(2):359–63.

David, A. and Sternberg, M. J. E. (2015). The contribution of missense mutations in core and rim residues of protein-protein interfaces to human disease. *Journal of molecular biology*, 427(17):2886–98.

de Beer, T. A. P., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., and Thornton, J. M. (2013). Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS computational biology*, 9(12):e1003382.

de Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., and Marrink, S. J. (2013). Improved parameters for the martini coarse-grained protein force field. *Journal of chemical theory and computation*, 9(1):687–97.

Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics (Oxford, England)*, 25(19):2537–43.

Django Software Foundation (2017). Django. http://djangoproject.com.

Donaldson, P., Daly, A., Ermini, L., and Bevitt, D. (2016). *Genetics of Complex Disease*. Garland Science.

Doss, C. G. P. and Nagasundaram, N. (2012). Investigating the structural impacts of I64T and P311S mutations in APE1-DNA complex: a molecular dynamics approach. *PloS one*, 7(2):e31677.

Drummond, D. A. and Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52.

Dunker, A. K., Bondos, S. E., Huang, F., and Oldfield, C. J. (2015). Intrinsically disordered proteins and multicellular organisms. *Seminars in cell and developmental biology*, 37:44–55.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440.

Durrant, J. D. and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology*, 9:71.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5:113.

Engin, H. B., Kreisberg, J. F., and Carter, H. (2016). Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PloS one*, 11(4):e0152929.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593.

Evilä, A., Palmio, J., Vihola, A., Savarese, M., Tasca, G., Penttilä, S., Lehtinen, S., Jonson, P. H., De Bleecker, J., Rainer, P., Auer-Grumbach, M., Pouget, J., Salort-Campana, E., Vilchez, J. J., Muelas, N., Olive, M., Hackman, P., and Udd, B. (2017). Targeted next-generation sequencing reveals novel TTN mutations causing recessive distal titinopathy. *Molecular neurobiology*, 54(9):7212–7223.

Evilä, A., Vihola, A., Sarparanta, J., Raheem, O., Palmio, J., Sandell, S., Eymard, B., Illa, I., Rojas-Garcia, R., Hankiewicz, K., Negrão, L., Löppönen, T., Nokelainen, P., Kärppä, M., Penttilä, S., Screen, M., Suominen, T., Richard, I., Hackman, P., and Udd, B. (2014). Atypical phenotypes in titinopathies explained by second titin mutations. *Annals of neurology*, 75(2):230–40.

Feder, J. N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D. A., Basava, A., Dormishian, F., Domingo, R., Ellis, M. C., Fullan, A., Hinton, L. M., Jones, N. L., Kimmel, B. E., Kronmal, G. S., Lauer, P., Lee, V. K., Loeb, D. B., Mapa, F. A., McClelland, E., Meyer, N. C., Mintier, G. A., Moeller, N., Moore, T., Morikang, E., Prass, C. E., Quintana, L., Starnes, S. M., Schatzman, R. C., Brunke, K. J., Drayna, D. T., Risch, N. J., Bacon, B. R., and Wolff, R. K. (1996). A novel MHC class i-like gene is mutated in patients with hereditary haemochromatosis. *Nature genetics*, 13(4):399–408.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y., and Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic acids research*, 45(D1):D190–D199.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). Pfam: the protein families database. *Nucleic acids research*, 42(Database issue):D222–30.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–85.

Follis, A. V., Llambi, F., Ou, L., Baran, K., Green, D. R., and Kriwacki, R. W. (2014). The DNA-binding domain mediates both nuclear and cytosolic functions of p53. *Nature structural and molecular biology*, 21(6):535–43.

Fong, J. H. and Panchenko, A. R. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Molecular bioSystems*, 6(10):1821–8.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(Database issue):D805–11.

Fornili, A., Pandini, A., Lu, H.-C., and Fraternali, F. (2013). Specialized dynamical properties of promiscuous residues revealed by simulated conformational ensembles. *Journal of chemical theory and computation*, 9(11):5127–5147.

Franken, H., Mathieson, T., Childs, D., Sweetman, G. M. A., Werner, T., Tögel, I., Doce, C., Gade, S., Bantscheff, M., Drewes, G., Reinhard, F. B. M., Huber, W., and Savitski, M. M. (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nature protocols*, 10(10):1567–93.

Fukuzawa, A., Lange, S., Holt, M., Vihola, A., Carmignac, V., Ferreiro, A., Udd, B., and Gautel, M. (2008). Interactions with titin and myomesin target obscurin and obscurin-like 1 to the M-band: implications for hereditary myopathies. *Journal of cell science*, 121(11):1841–51.

Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M., and Orengo, C. A. (2016). Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. *Journal of molecular biology*, 428(2 Pt A):253–267.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science (New York, N.Y.)*, 296(5576):2225–9.

Ganna, A., Genovese, G., Howrigan, D. P., Byrnes, A., Kurki, M., Zekavat, S. M., Whelan, C. W., Kals, M., Nivard, M. G., Bloemendal, A., Bloom, J. M., Goldstein, J. I., Poterba, T., Seed, C., Handsaker, R. E., Natarajan, P., Mägi, R., Gage, D., Robinson, E. B., Metspalu, A., Salomaa, V., Suvisaari, J., Purcell, S. M., Sklar, P., Kathiresan, S., Daly, M. J., McCarroll, S. A., Sullivan, P. F., Palotie, A., Esko, T., Hultman, C., and Neale, B. M. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nature neuroscience*, 19(12):1563–1565.

Gao, M., Zhou, H., and Skolnick, J. (2015). Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure (London, England : 1993)*, 23(7):1362–9.

Garcia, T. I., Oberhauser, A. F., and Braun, W. (2009). Mechanical stability and differentially conserved physical-chemical properties of titin Ig-domains. *Proteins*, 75(3):706–18.

Garg, M., Braunstein, G., and Koeffler, H. P. (2014). LAMC2 as a therapeutic target for cancers. *Expert opinion on therapeutic targets*, 18(9):979–82.

Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., Smits, G., and Lenaerts, T. (2016). DIDA: A curated and annotated digenic diseases database. *Nucleic acids research*, 44(D1):D900–7.

Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitás, K., Sasse-Klaassen, S., Seidman, J. G., Seidman, C., Granzier, H., Labeit, S., Frenneaux, M., and Thierfelder, L. (2002). Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nature genetics*, 30(2):201–4.

Gibbs, E. B. and Kriwacki, R. W. (2018). Direct detection of carbon and nitrogen nuclei for high-resolution analysis of intrinsically disordered proteins using NMR spectroscopy. *Methods*, 138-139:39–46.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature reviews. Genetics*, 13(2):135–45.

Gigli, M., Begay, R. L., Morea, G., Graw, S. L., Sinagra, G., Taylor, M. R. G., Granzier, H., and Mestroni, L. (2016). A review of the giant protein titin in clinical molecular diagnostics of cardiomyopathies. *Frontiers in cardiovascular medicine*, 3:21.

Goldstein, D. B. and Cavalleri, G. L. (2005). Genomics: understanding human diversity. *Nature*, 437(7063):1241–2.

Gommans, W. M., Mullen, S. P., and Maas, S. (2009). RNA editing: a driving force for adaptive evolution? *BioEssays : news and reviews in molecular, cellular and developmental biology*, 31(10):1137–45.

Gong, S. and Blundell, T. L. (2010). Structural and functional restraints on the occurrence of single amino acid variations in human proteins. *PloS one*, 5(2):e9186.

González-Pérez, A. and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, Condel. *American journal of human genetics*, 88(4):440–9.

Gorina, S. and Pavletich, N. P. (1996). Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science (New York, N.Y.)*, 274(5289):1001–5.

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)*, 22(21):2695–6.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, 185(4154):862–4.

Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J., and Fowler, D. M. (2018). Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell systems*, 6(1):116–124.e3.

Gress, A., Ramensky, V., and Kalinina, O. V. (2017). Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes. *Oncogenesis*, 6(9):e380.

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., and Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human mutation*, 36(5):513–23.

Gromiha, M. M. and Sarai, A. (2010). Thermodynamic database for proteins: features and applications. *Methods in molecular biology (Clifton, N.J.)*, 609:97–112.

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580–5.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–87.

Hackman, P., Vihola, A., Haravuori, H., Marchand, S., Sarparanta, J., De Seze, J., Labeit, S., Witt, C., Peltonen, L., Richard, I., and Udd, B. (2002). Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. *American journal of human genetics*, 71(3):492–500.

Hagberg, A., Swart, P., and S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hastings, R., de Villiers, C. P., Hooper, C., Ormondroyd, L., Pagnamenta, A., Lise, S., Salatino, S., Knight, S. J. L., Taylor, J. C., Thomson, K. L., Arnold, L., Chatziefthimiou, S. D., Konarev, P. V., Wilmanns, M., Ehler, E., Ghisleni, A., Gautel, M., Blair, E., Watkins, H., and Gehmlich, K. (2016). Combination of whole genome sequencing, linkage, and functional studies implicates a missense mutation in titin as a cause of autosomal dominant cardiomyopathy with features of left ventricular noncompaction. *Circulation. Cardiovascular genetics*, 9(5):426–435.

Hauser, A. S., Chavali, S., Masuho, I., Jahn, L. J., Martemyanov, K. A., Gloriam, D. E., and Babu, M. M. (2018). Pharmacogenomics of GPCR drug targets. *Cell*, 172(1-2):41–54.e19.

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., un Nisa Dar, V., Bezginov, A., Clark, G. W., Wu, G. C., Wodak, S. J., Tillier, E. R. M., Paccanaro, A., Marcotte, E. M., and Emili, A. (2012). A census of human soluble protein complexes. *Cell*, 150(5):1068–81.

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC genomics*, 16 Suppl 8:S1.

Hedberg, C., Toledo, A. G., Gustafsson, C. M., Larson, G., Oldfors, A., and Macao, B. (2014). Hereditary myopathy with early respiratory failure is associated with misfolding of the titin fibronectin III 119 subdomain. *Neuromuscular disorders : NMD*, 24(5):373–9.

Helle, E. and Parikh, V. N. (2016). Wrestling the giant: New approaches for assessing titin variant pathogenicity. *Circulation. Cardiovascular genetics*, 9(5):392–394.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–9.

Herman, D. S., Lam, L., Taylor, M. R. G., Wang, L., Teekakirikul, P., Christodoulou, D., Conner, L., DePalma, S. R., McDonough, B., Sparks, E., Teodorescu, D. L., Cirino, A. L., Banner, N. R., Pennell, D. J., Graw, S., Merlo, M., Lenarda, A. D., Sinagra, G., Bos, J. M., Ackerman, M. J., Mitchell, R. N., Murry, C. E., Lakdawala, N. K., Ho, C. Y., Barton, P. J. R., Cook, S. A., Mestroni, L., Seidman, J. G., and Seidman, C. E. (2012). Truncations of titin causing dilated cardiomyopathy. *The New England journal of medicine*, 366(7):619–28.

Hess, B. (2008). P-LINCS: a parallel linear constraint solver for molecular simulation. *Journal of chemical theory and computation*, 4(1):116–22.

Holehouse, A. S. and Naegle, K. M. (2015). Reproducible analysis of post-translational modifications in proteomes–application to human mutations. *PloS one*, 10(12):e0144692.

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome medicine*, 5(9):79.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135.

Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–61.

Hospital, A., Goñi, J. R., Orozco, M., and Gelpí, J. L. (2015). Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry : AABC*, 8:37–47.

Hou, J. P. and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. *Genome medicine*, 6(7):56.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Hurst, J. M., McMillan, L. E. M., Porter, C. T., Allen, J., Fakorede, A., and Martin, A. C. R. (2009). The SAAPdb web resource: a large-scale structural analysis of mutant proteins. *Human mutation*, 30(4):616–24.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., Kolippakkam, D., Mintseris, J., Obar, R. A., Harris, T., Artavanis-Tsakonas, S., Sowa, M. E., Camilli, P. D., Paulo, J. A., Harper, J. W., and Gygi, S. P. (2015). The BioPlex network: A systematic exploration of the human interactome. *Cell*, 162(2):425–440.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45.

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., Powell, I. J., Cussenot, O., Cancel-Tassin, G., Giles, G. G., MacInnis, R. J., Maier, C., Hsieh, C.-L., Wiklund, F., Catalona, W. J., Foulkes, W. D., Mandal, D., Eeles, R. A., Kote-Jarai, Z., Bustamante, C. D., Schaid, D. J., Hastie, T., Ostrander, E. A., Bailey-Wilson, J. E., Radivojac, P., Thibodeau, S. N., Whittemore, A. S., and Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American journal of human genetics*, 99(4):877–885.

Itoh-Satoh, M., Hayashi, T., Nishi, H., Koga, Y., Arimura, T., Koyanagi, T., Takahashi, M., Hohda, S., Ueda, K., Nouchi, T., Hiroe, M., Marumo, F., Imaizumi, T., Yasunami, M., and Kimura, A. (2002). Titin mutations as the molecular basis for dilated cardiomyopathy. *Biochemical and biophysical research communications*, 291(2):385–93.

Izumi, R., Niihori, T., Aoki, Y., Suzuki, N., Kato, M., Warita, H., Takahashi, T., Tateyama, M., Nagashima, T., Funayama, R., Abe, K., Nakayama, K., Aoki, M., and Matsubara, Y. (2013). Exome sequencing identifies a novel TTN mutation in a family with hereditary myopathy with early respiratory failure. *Journal of human genetics*, 58(5):259–66.

Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J., and Moal, I. H. (2018). SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics (Oxford, England)*.

Jeanes, A., Gottardi, C. J., and Yap, A. S. (2008). Cadherins and cancer: how does cadherin dysfunction promote tumor progression? *Oncogene*, 27(55):6920–6929.

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., and Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5:4022.

Jones, D. T. and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics (Oxford, England)*, 31(6):857–63.

Jubb, H., Blundell, T. L., and Ascher, D. B. (2015). Flexibility and small pockets at protein-protein interfaces: New insights into druggability. *Progress in biophysics and molecular biology*, 119(1):2–9.

Jubb, H. C., Pandurangan, A. P., Turner, M. A., Ochoa-Montaño, B., Blundell, T. L., and Ascher, D. B. (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*, 128:3–13.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.-P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J., and Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature genetics*, 47(7):818–21.

Kawabata, T., Ota, M., and Nishikawa, K. (1999). The protein mutant database. *Nucleic acids research*, 27(1):355–7.

Kellermayer, M., Sziklai, D., Papp, Z., Decker, B., Lakatos, E., and Mártonfalvi, Z. (2018). Topology of interaction between titin and myosin thick filaments. *Journal of structural biology*, 203(1):46–53.

Kellogg, E. H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, 79(3):830–8.

Kenny, P. A., Liston, E. M., and Higgins, D. G. (1999). Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene*, 232(1):11–23.

Kenzaki, H., Koga, N., Hori, N., Kanada, R., Li, W., Okazaki, K.-I., Yao, X.-Q., and Takada, S. (2011). CafeMol: A coarse-grained biomolecular simulator for simulating proteins at work. *Journal of chemical theory and computation*, 7(6):1979–89.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55.

Kido, T., Sikora-Wohlfeld, W., Kawashima, M., Kikuchi, S., Kamatani, N., Patwardhan, A., Chen, R., Sirota, M., Kodama, K., Hadley, D., and Butte, A. J. (2018). Are minor alleles more likely to be risk alleles? *BMC medical genomics*, 11(1):3.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5.

Kitsios, G. D. and Zintzaras, E. (2009). Genome-wide association studies: hypothesis-"free" or "engaged"? *Translational research : the journal of laboratory and clinical medicine*, 154(4):161–4.

Kleinjung, J. and Fraternali, F. (2005). POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic acids research*, 33(Web Server issue):W342–6.

Kleinjung, J. and Fraternali, F. (2014). Design and application of implicit solvent models in biomolecular simulations. *Current opinion in structural biology*, 25:126–34.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45.

Kumar, A. and Purohit, R. (2014). Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS computational biology*, 10(4):e1003318.

Kumar, D. T., Doss, C. G. P., Sneha, P., Tayubi, I. A., Siva, R., Chakraborty, C., and Magesh, R. (2017). Influence of V54M mutation in giant muscle protein titin: a computational screening and molecular dynamics approach. *Journal of biomolecular structure and dynamics*, 35(5):917–928.

Kutzner, C., Páll, S., Fechner, M., Esztermann, A., de Groot, B. L., and Grubmüller, H. (2015). Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of computational chemistry*, 36(26):1990–2008.

Laddach, A., Gautel, M., and Fraternali, F. (2017). TITINdb-a computational tool to assess titin's role as a disease gene. *Bioinformatics (Oxford, England)*, 33(21):3482–3485.

Laddach, A., Ng, J. C.-F., Chung, S. S., and Fraternali, F. (2018). Genetic variants and protein-protein interactions: a multidimensional network-centric view. *Current opinion in structural biology*, 50:82–90.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–8.

Laskowski, R. A. and Thornton, J. M. (2008). Understanding the molecular machinery of genetics through 3D structures. *Nature reviews. Genetics*, 9(2):141–51.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., and MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–91.

Leontyev, I. V. and Stuchebrukhov, A. A. (2010). Electronic polarizability and the effective pair potentials of water. *Journal of chemical theory and computation*, 6(10):3153–3161.

Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P. J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science (New York, N.Y.)*, 355(6327).

Levy, E. D., De, S., and Teichmann, S. A. (2012). Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50):20461–6.

Lewis, C. M. and Knight, J. (2012). Introduction to genetic association studies. *Cold Spring Harbor protocols*, 2012(3):297–306.

Lezon, T. R., Shrivastava, I. H., Yang, Z., and Bahar, I. (2010). Elastic network models for biomolecular dynamics: theory and application to membrane proteins and viruses. In *Handbook on Biological Networks*, pages 129–158. World Scientific.

Li, M.-X., Kwan, J. S. H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., and Sham, P. C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS genetics*, 9(1):e1003143.

Li, Y. and Linke, W. A. (2017). Mechanically unfolded titin immunoglobulin domains refold faster and more accurately in presence of chaperone alpha-B-crystallin. *Biophysical Journal*, 112(3):42a.

LIU, X., RAO, L., ZHOU, B., lei ZHANG, B., yun WANG, Y., CHEN, B., WU, Y., and HUANG, P. (2008). [titin gene mutations in chinese patients with dilated cardiomyopathy]. *Zhonghua xin xue guan bing za zhi*, 36(12):1066–9.

Lopes, L. R., Zekavati, A., Syrris, P., Hubank, M., Giambartolomei, C., Dalageorgou, C., Jenkins, S., McKenna, W., Plagnol, V., and Elliott, P. M. (2013). Genetic complexity in hypertrophic cardiomyopathy revealed by high-throughput sequencing. *Journal of medical genetics*, 50(4):228–39.

Lu, C., Jain, S. U., Hoelper, D., Bechet, D., Molden, R. C., Ran, L., Murphy, D., Venneti, S., Hameed, M., Pawel, B. R., Wunder, J. S., Dickson, B. C., Lundgren, S. M., Jani, K. S., De Jay, N., Papillon-Cavanagh, S., Andrulis, I. L., Sawyer, S. L., Grynspan, D., Turcotte, R. E., Nadaf, J., Fahiminiyah, S., Muir, T. W., Majewski, J., Thompson, C. B., Chi, P., Garcia, B. A., Allis, C. D., Jabado, N., and Lewis, P. W. (2016a). Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science (New York, N.Y.)*, 352(6287):844–9.

Lu, H.-C., Braga, J. H., and Fraternali, F. (2016b). PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks. *Bioinformatics (Oxford, England)*, 32(16):2534–6.

Lu, H.-C., Chung, S. S., Fornili, A., and Fraternali, F. (2015). Anatomy of protein disorder, flexibility and disease-related mutations. *Frontiers in molecular biosciences*, 2:47.

Luscombe, N. M. and Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of molecular biology*, 320(5):991–1009.

Luzzatto, L. (2012). Sickle cell anaemia and malaria. *Mediterranean journal of hematology and infectious diseases*, 4(1):e2012065.

Mahlich, Y., Reeb, J., Hecht, M., Schelling, M., De Beer, T. A. P., Bromberg, Y., and Rost, B. (2017). Common sequence variants affect molecular function more than rare variants? *Scientific reports*, 7(1):1608.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS letters*, 583(24):3966–73.

Makwana, K. M. and Mahalakshmi, R. (2015). Implications of aromatic-aromatic interactions: From protein structures to peptide models. *Protein science : a publication of the Protein Society*, 24(12):1920–33.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.

Maron, B. J., Gardin, J. M., Flack, J. M., Gidding, S. S., Kurosaki, T. T., and Bild, D. E. (1995). Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA study. Coronary Artery Risk Development in (Young) Adults. *Circulation*, 92(4):785–9.

Mathieson, T., Franken, H., Kosinski, J., Kurzawa, N., Zinn, N., Sweetman, G., Poeckel, D., Ratnu, V. S., Schramm, M., Becher, I., Steidel, M., Noh, K.-M., Bergamini, G., Beck, M., Bantscheff, M., and Savitski, M. M. (2018). Systematic analysis of protein turnover in primary cells. *Nature communications*, 9(1):689.

Matsumoto, Y., Hayashi, T., Inagaki, N., Takahashi, M., Hiroi, S., Nakamura, T., Arimura, T., Nakamura, K., Ashizawa, N., Yasunami, M., Ohe, T., Yano, K., and Kimura, A. (2005). Functional analysis of titin/connectin N2-B mutations found in cardiomyopathy. *Journal of muscle research and cell motility*, 26(6-8):367–74.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology*, 17(1):122.

Mehta, V. and Sharma, C. V. (2013). Paracetamol: mechanisms and updates. *Continuing Education in Anaesthesia Critical Care & Pain*, 14(4):153–158.

Meier, K., Choutko, A., Dolenc, J., Eichenberger, A. P., Riniker, S., and van Gunsteren, W. F. (2013). Multi-resolution simulation of biomolecular systems: a review of methodological issues. *Angewandte Chemie (International ed. in English)*, 52(10):2820–34.

Meyer, L. C. and Wright, N. T. (2013). Structure of giant muscle proteins. *Frontiers in physiology*, 4:368.

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B., and Beckstein, O. (2011). MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, 32(10):2319–27.

Miller, M., Bromberg, Y., and Swint-Kruse, L. (2017). Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Scientific reports*, 7:41329.

Miosge, L. A., Field, M. A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S., Beutler, B., Whittle, B., Bertram, E. M., Enders, A., Goodnow, C. C., and Andrews, T. D. (2015). Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37):E5189–98.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The MARTINI coarse-grained force field: Extension to proteins. *Journal of chemical theory and computation*, 4(5):819–34.

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., and Aloy, P. (2015). dSysMap: exploring the edgetic role of disease mutations. *Nature methods*, 12(3):167–8.

Mottaz, A., David, F. P. A., Veuthey, A.-L., and Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics (Oxford, England)*, 26(6):851–2.

Muller, P. A. J. and Vousden, K. H. (2013). p53 mutations in cancer. *Nature cell biology*, 15(1):2–8.

Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Computer applications in the biosciences : CABIOS*, 4(1):11–7.

MySQL, A. (2008). *Mysql 5.1 reference manual*.

Nair, P. S. and Vihinen, M. (2013). VariBench: a benchmark database for variations. *Human mutation*, 34(1):42–9.

Ng, P. C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–74.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A. R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PloS one*, 8(6):e66273.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17.

Ohlsson, M., Hedberg, C., Brådvik, B., Lindberg, C., Tajsharghi, H., Danielsson, O., Melberg, A., Udd, B., Martinsson, T., and Oldfors, A. (2012). Hereditary myopathy with early respiratory failure associated with a mutation in A-band titin. *Brain : a journal of neurology*, 135(Pt 6):1682–94.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–45.

Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

Oliver, G. R., Zimmermann, M. T., Klee, E. W., and Urrutia, R. A. (2016). "the molecule's the thing:" the promise of molecular modeling and dynamic simulations in aiding the prioritization and interpretation of genomic testing results. *F1000Research*, 5:766.

Olow, A., Chen, Z., Niedner, R. H., Wolf, D. M., Yau, C., Pankov, A., Lee, E. P. R., Brown-Swigart, L., van 't Veer, L. J., and Coppé, J.-P. (2016). An atlas of the human kinome reveals the mutational landscape underlying dysregulated phosphorylation cascades in cancer. *Cancer research*, 76(7):1733–45.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(Database issue):D358–63.

Orlov, I., Myasnikov, A. G., Andronov, L., Natchiar, S. K., Khatter, H., Beinsteiner, B., Ménétret, J.-F., Hazemann, I., Mohideen, K., Tazibt, K., Tabaroni, R., Kratzat, H., Djabeur, N., Bruxelles, T., Raivoniaina, F., di Pompeo, L., Torchy, M., Billas, I., Urzhumtsev, A., and Klaholz, B. P. (2017). The integrative role of cryo electron microscopy in molecular and cellular structural biology. *Biology of the cell*, 109(2):81–93.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of molecular biology*, 340(2):385–95.

Ott, J., Wang, J., and Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nature reviews. Genetics*, 16(5):275–84.

Palmio, J., Evilä, A., Chapon, F., Tasca, G., Xiang, F., Brådvik, B., Eymard, B., Echaniz-Laguna, A., Laporte, J., Kärppä, M., Mahjneh, I., Quinlivan, R., Laforêt, P., Damian, M., Berardo, A., Taratuto, A. L., Bueri, J. A., Tommiska, J., Raivio, T., Tuerk, M., Gölitz, P., Chevessier, F., Sewry, C., Norwood, F., Hedberg, C., Schröder, R., Edström, L., Oldfors, A., Hackman, P., and Udd, B. (2014). Hereditary myopathy with early respiratory failure: occurrence in various populations. *Journal of neurology, neurosurgery, and psychiatry*, 85(3):345–53.

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2012). Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 26(2):868–81.

Pandini, A., Fornili, A., Fraternali, F., and Kleinjung, J. (2013). GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. *Bioinformatics (Oxford, England)*, 29(16):2053–5.

Pandurangan, A. P., Ochoa-Montaño, B., Ascher, D. B., and Blundell, T. L. (2017). SDM: a server for predicting effects of mutations on protein stability. *Nucleic acids research*, 45(W1):W229–W235.

Pantazis, A., Vischer, A. S., Perez-Tome, M. C., and Castelletti, S. (2015). Diagnosis and management of hypertrophic cardiomyopathy. *Echo research and practice*, 2(1):R45–53.

Parrinello, M. and Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190.

Patnala, R., Clements, J., and Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC genetics*, 14:39.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K. B., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S., and Pandey, A. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, 32(Database issue):D497–501.

Periole, X., Cavalli, M., Marrink, S.-J., and Ceruso, M. A. (2009). Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition. *Journal of chemical theory and computation*, 5(9):2531–43.

Pernigo, S., Fukuzawa, A., Pandini, A., Holt, M., Kleinjung, J., Gautel, M., and Steiner, R. A. (2015). The crystal structure of the human titin:obscurin complex reveals a conserved yet specific muscle M-band zipper module. *Journal of molecular biology*, 427(4):718–736.

Petschnigg, J., Kotlyar, M., Blair, L., Jurisica, I., Stagljar, I., and Ketteler, R. (2017). Systematic identification of oncogenic EGFR interaction partners. *Journal of molecular biology*, 429(2):280–294.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612.

Petukh, M., Kucukkal, T. G., and Alexov, E. (2015). On human disease-causing amino acid variants: statistical study of sequence and structural patterns. *Human mutation*, 36(5):524–534.

PFAM (2018a). Family: NPIP (PF06409). https://pfam.xfam.org/family/PF06409. [Online; accessed 13-Mar-2018].

PFAM (2018b). Family: NUT (PF12881). https://pfam.xfam.org/family/PF12881. [Online; accessed 13-Mar-2018].

Pfeffer, G., Elliott, H. R., Griffin, H., Barresi, R., Miller, J., Marsh, J., Evilä, A., Vihola, A., Hackman, P., Straub, V., Dick, D. J., Horvath, R., Santibanez-Koref, M., Udd, B., and Chinnery, P. F. (2012). Titin mutation segregates with hereditary myopathy with early respiratory failure. *Brain : a journal of neurology*, 135(Pt 6):1695–713.

Pfeffer, G., Povitz, M., Gibson, G. J., and Chinnery, P. F. (2015). Diagnosis of muscle diseases presenting with early respiratory failure. *Journal of neurology*, 262(5):1101–14.

Pfeffer, G., Sambuughin, N., Olivé, M., Tyndel, F., Toro, C., Goldfarb, L. G., and Chinnery, P. F. (2014). A new disease allele for the p.C30071R mutation in titin causing hereditary myopathy with early respiratory failure. *Neuromuscular disorders : NMD*, 24(3):241–4.

Pieper, U., Eswar, N., Webb, B. M., Eramian, D., Kelly, L., Barkan, D. T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M. A., Davis, F. P., and Sali, A. (2009). MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 37(Database issue):D347–54.

Pires, A. S., Porto, W. F., Franco, O. L., and Alencar, S. A. (2017). In silico analyses of deleterious missense SNPs of human apolipoprotein E3. *Scientific reports*, 7(1):2509.

Pires, D. E. V. and Ascher, D. B. (2017). mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. *Nucleic acids research*, 45(W1):W241–W246.

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014a). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic acids research*, 42(Web Server issue):W314–9.

Pires, D. E. V., Ascher, D. B., and Blundell, T. L. (2014b). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics (Oxford, England)*, 30(3):335–42.

Pollazzon, M., Suominen, T., Penttilä, S., Malandrini, A., Carluccio, M. A., Mondelli, M., Marozza, A., Federico, A., Renieri, A., Hackman, P., Dotti, M. T., and Udd, B. (2010). The first Italian family with tibial muscular dystrophy caused by a novel titin mutation. *Journal of neurology*, 257(4):575–9.

Poma, A. B., Cieplak, M., and Theodorakis, P. E. (2017). Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *Journal of chemical theory and computation*, 13(3):1366–1374.

Ponzoni, L. and Bahar, I. (2018). Structural dynamics is a determinant of the functional significance of missense variants. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):4164–4169.

Porollo, A. and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. *Proteins*, 66(3):630–45.

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015a). A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLoS computational biology*, 11(10):e1004518.

Porta-Pardo, E. and Godzik, A. (2014). e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics (Oxford, England)*, 30(21):3109–14.

Porta-Pardo, E., Hrabe, T., and Godzik, A. (2015b). Cancer3D: understanding cancer mutations through protein structures. *Nucleic acids research*, 43(Database issue):D968–73.

Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G., and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nature methods*, 14(8):782–788.

Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C.-H., Lu, Z., Boutet, E., Bye-A-Jee, H., Famiglietti, M. L., Roechert, B., and Consortium, T. U. (2017). On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics (Oxford, England)*, 33(21):3454–3460.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (Refseq): current status, new features and genome annotation policy. *Nucleic acids research*, 40(Database issue):D130–5.

Pucci, F., Bourgeas, R., and Rooman, M. (2016). Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific reports*, 6:23257.

Pucci, F. and Rooman, M. (2016). Improved insights into protein thermal stability: from the molecular to the structurome scale. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2080).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–75.

Quintana-Murci, L. (2016). Understanding rare and common diseases in the context of human evolution. *Genome biology*, 17(1):225.

Raza, S. and Hall, A. (2017). Genomic medicine and data sharing. *British medical bulletin*, 123(1):35–45.

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in genetics : TIG*, 17(9):502–10.

Reif, M. M., Winger, M., and Oostenbrink, C. (2013). Testing of the GROMOS force-field parameter set 54A8: Structural properties of electrolyte solutions, lipid bilayers, and proteins. *Journal of chemical theory and computation*, 9(2):1247–1264.

Reimand, J., Wagih, O., and Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific reports*, 3:2651.

Reimand, J., Wagih, O., and Bader, G. D. (2015). Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS genetics*, 11(1):e1004919.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118.

Rief, M., Gautel, M., Schemmel, A., and Gaub, H. E. (1998). The mechanical stability of immunoglobulin and fibronectin III domains in the muscle protein titin measured by atomic force microscopy. *Biophysical journal*, 75(6):3008–14.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annual review of biochemistry*, 79:233–69.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A.-R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., MacWilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J.-C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A.-L., Iakoucheva, L. M., Aloy, P., Rivas, J. D. L., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P., and Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Roncarati, R., Anselmi, C. V., Krawitz, P., Lattanzi, G., von Kodolitsch, Y., Perrot, A., di Pasquale, E., Papa, L., Portararo, P., Columbaro, M., Forni, A., Faggian, G., Condorelli, G., and Robinson, P. N. (2013). Doubly heterozygous LMNA and TTN mutations revealed by exome sequencing in a severe form of dilated cardiomyopathy. *European journal of human genetics : EJHG*, 21(10):1105–11.

Rudloff, M. W., Woosley, A. N., and Wright, N. T. (2015). Biophysical characterization of naturally occurring titin M10 mutations. *Protein science : a publication of the Protein Society*, 24(6):946–55.

Sahni, N., Yi, S., Taipale, M., Bass, J. I. F., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y.-Y., Yachie, N., Zhong, Q., Shen, Y., Palagi, A., San-Miguel, A., Fan, C., Balcha, D., Dricot, A., Jordan, D. M., Walsh, J. M., Shah, A. A., Yang, X., Stoyanova, A. K., Leighton, A., Calderwood, M. A., Jacob, Y., Cusick, M. E., Salehi-Ashtiani, K., Whitesell, L. J., Sunyaev, S., Berger, B., Barabási, A.-L., Charloteaux, B., Hill, D. E., Hao, T., Roth, F. P., Xia, Y., Walhout, A. J. M., Lindquist, S., and Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3):647–660.

Sakakura, M., Hadziselimovic, A., Wang, Z., Schey, K. L., and Sanders, C. R. (2011). Structural basis for the Trembler-J phenotype of Charcot-Marie-Tooth disease. *Structure (London, England : 1993)*, 19(8):1160–9.

Sánchez, R. and Sali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13597–602.

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., and Overington, J. P. (2017). A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery*, 16(1):19–34.

Satoh, M., Takahashi, M., Sakamoto, T., Hiroe, M., Marumo, F., and Kimura, A. (1999). Structural analysis of the titin gene in hypertrophic cardiomyopathy: identification of a novel disease gene. *Biochemical and biophysical research communications*, 262(2):411–7.

Savarese, M., Maggi, L., Vihola, A., Jonson, P. H., Tasca, G., Ruggiero, L., Bello, L., Magri, F., Giugliano, T., Torella, A., Evilä, A., Fruscio, G. D., Vanakker, O., Gibertini, S., Vercelli, L., Ruggieri, A., Antozzi, C., Luque, H., Janssens, S., Pasanisi, M. B., Fiorillo, C., Raimondi, M., Ergoli, M., Politano, L., Bruno, C., Rubegni, A., Pane, M., Santorelli, F. M., Minetti, C., Angelini, C., De Bleecker, J., Moggio, M., Mongini, T., Comi, G. P., Santoro, L., Mercuri, E., Pegoraro, E., Mora, M., Hackman, P., Udd, B., and Nigro, V. (2018). Interpreting genetic variants in titin in patients with muscle disorders. *JAMA neurology*, 75(5):557–565.

Savarese, M., Sarparanta, J., Vihola, A., Udd, B., and Hackman, P. (2016). Increasing role of titin mutations in neuromuscular disorders. *Journal of neuromuscular diseases*, 3(3):293–308.

Sazonovs, A. and Barrett, J. C. (2018). Rare-variant studies to complement genome-wide association studies. *Annual review of genomics and human genetics*, 19:97–112.

Schafer, S., de Marvao, A., Adami, E., Fiedler, L. R., Ng, B., Khin, E., Rackham, O. J. L., van Heesch, S., Pua, C. J., Kui, M., Walsh, R., Tayal, U., Prasad, S. K., Dawes, T. J. W., Ko, N. S. J., Sim, D., Chan, L. L. H., Chin, C. W. L., Mazzarotto, F., Barton, P. J., Kreuchwig, F., de Kleijn, D. P. V., Totman, T., Biffi, C., Tee, N., Rueckert, D., Schneider, V., Faber, A., Regitz-Zagrosek, V., Seidman, J. G., Seidman, C. E., Linke, W. A., Kovalik, J.-P., O'Regan, D., Ware, J. S., Hubner, N., and Cook, S. A. (2017). Titin-truncating variants affect heart function in disease cohorts and the general population. *Nature genetics*, 49(1):46–53.

Schaffner, M. and Benini, L. (2018). On the feasibility of FPGA acceleration of molecular dynamics simulations. *arXiv preprint arXiv:1808.04201*.

Schneider, B., Cernỳ, J., Svozil, D., Cech, P., Gelly, J.-C., and de Brevern, A. G. (2014). Bioinformatic analysis of the protein/DNA interface. *Nucleic acids research*, 42(5):3381–94.

Schurmann, K., Anton, M., Ivanov, I., Richter, C., Kuhn, H., and Walther, M. (2011). Molecular basis for the reduced catalytic activity of the naturally occurring T560M mutant of human 12/15-lipoxygenase that has been implicated in coronary artery disease. *The Journal of biological chemistry*, 286(27):23920–7.

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*, 7(8):575–6.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic acids research*, 33(Web Server issue):W382–8.

Seeliger, D. and de Groot, B. L. (2010). Protein thermostability calculations using alchemical free energy simulations. *Biophysical journal*, 98(10):2309–16.

Seidman, C. E. and Seidman, J. G. (2011). Identifying sarcomere gene mutations in hypertrophic cardiomyopathy: a personal history. *Circulation research*, 108(6):743–50.

Semenza, G. L. (2006). VHL and p53: tumor suppressors team up to prevent cancer. *Molecular cell*, 22(4):437–9.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*.

Serohijos, A. W. R., Rimas, Z., and Shakhnovich, E. I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports*, 2(2):249–56.

Shen, M.-Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein science : a publication of the Protein Society*, 15(11):2507–24.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–11.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., and Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, 34(1):57–65.

Signorell, A. (2017). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.19.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(Database issue):D376–81.

Sivley, R. M., Dou, X., Meiler, J., Bush, W. S., and Capra, J. A. (2018). Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *American journal of human genetics*, 102(3):415–426.

Stehr, H., Jang, S.-H. J., Duarte, J. M., Wierling, C., Lehrach, H., Lappe, M., and Lange, B. M. H. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Molecular cancer*, 10:54.

Steinbrecher, T., Zhu, C., Wang, L., Abel, R., Negron, C., Pearlman, D., Feyfant, E., Duan, J., and Sherman, W. (2017). Predicting the effect of amino acid single-point mutations on protein stability-large-scale validation of MD-based relative free energy calculations. *Journal of molecular biology*, 429(7):948–963.

Stone, E. A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome research*, 15(7):978–86.

Strom, C. M., Crossley, B., Buller-Buerkle, A., Jarvis, M., Quan, F., Peng, M., Muralidharan, K., Pratt, V., Redman, J. B., and Sun, W. (2011). Cystic fibrosis testing 8 years on: lessons learned from carrier screening and sequencing analysis. *Genetics in medicine : official journal of the American College of Medical Genetics*, 13(2):166–72.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50.

Sunyaev, S., Ramensky, V., and Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, 16(5):198–200.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., and Bork, P. (2001). Prediction of deleterious human alleles. *Human molecular genetics*, 10(6):591–7.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., and Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein engineering*, 12(5):387–94.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(Database issue):D447–52.

Takada, S., Kanada, R., Tan, C., Terakawa, T., Li, W., and Kenzaki, H. (2015). Modeling structural dynamics of biomolecular complexes by coarse-grained molecular simulations. *Accounts of chemical research*, 48(12):3026–35.

Takano, K., Liu, D., Tarpey, P., Gallant, E., Lam, A., Witham, S., Alexov, E., Chaubey, A., Stevenson, R. E., Schwartz, C. E., Board, P. G., and Dulhunty, A. F. (2012). An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity. *Human molecular genetics*, 21(20):4497–507.

Tang, H. and Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, 203(2):635–47.

The UniProt Consortium (2018). UniProt: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169.

Theobald, D. L. and Wuttke, D. S. (2006). THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics (Oxford, England)*, 22(17):2171–2.

Thomas, P. D. and Kejariwal, A. (2004). Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15398–403.

Tironi, I. G., Sperb, R., Smith, P. E., and van Gunsteren, W. F. (1995). A generalized reaction field method for molecular dynamics simulations. *The Journal of Chemical Physics*, 102(13):5451–5459.

Tomita, Y., Marchenko, N., Erster, S., Nemajerova, A., Dehner, A., Klein, C., Pan, H., Kessler, H., Pancoska, P., and Moll, U. M. (2006). WT p53, but not tumor-derived mutants, bind to Bcl2 via the dna binding domain and induce mitochondrial permeabilization. *The Journal of biological chemistry*, 281(13):8600–6.

Topham, C. M., Srinivasan, N., and Blundell, T. L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein engineering*, 10(1):7–21.

Toro, C., Olivé, M., Dalakas, M. C., Sivakumar, K., Bilbao, J. M., Tyndel, F., Vidal, N., Farrero, E., Sambuughin, N., and Goldfarb, L. G. (2013). Exome sequencing identifies titin mutations causing hereditary myopathy with early respiratory failure (HMERF) in families of diverse ethnic origins. *BMC neurology*, 13:29.

Tsao, H. and Florez, J. C. (2007). Introduction to genetic association studies. *The Journal of investigative dermatology*, 127(10):2283–7.

Uruha, A., Hayashi, Y. K., Oya, Y., Mori-Yoshimura, M., Kanai, M., Murata, M., Kawamura, M., Ogata, K., Matsumura, T., Suzuki, S., Takahashi, Y., Kondo, T., Kawarabayashi, T., Ishii, Y., Kokubun, N., Yokoi, S., Yasuda, R., ichi Kira, J., Mitsuhashi, S., Noguchi, S., Nonaka, I., and Nishino, I. (2015). Necklace cytoplasmic bodies in hereditary myopathy with early respiratory failure. *Journal of neurology, neurosurgery, and psychiatry*, 86(5):483–9.

Van den Bergh, P. Y. K., Bouquiaux, O., Verellen, C., Marchand, S., Richard, I., Hackman, P., and Udd, B. (2003). Tibial muscular dystrophy in a Belgian family. *Annals of neurology*, 54(2):248–51.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–18.

van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glättli, A., Hünenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F. A., and Yu, H. B. (2006). Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie (International ed. in English)*, 45(25):4064–92.

van Gunsteren, W. F., Daura, X., Hansen, N., Mark, A. E., Oostenbrink, C., Riniker, S., and Smith, L. J. (2018). Validation of molecular simulation: An overview of issues. *Angewandte Chemie (International ed. in English)*, 57(4):884–902.

van Spaendonck-Zwarts, K. Y., Posafalvi, A., van den Berg, M. P., Hilfiker-Kleiner, D., Bollen, I. A. E., Sliwa, K., Alders, M., Almomani, R., van Langen, I. M., van der Meer, P., Sinke, R. J., van der Velden, J., Van Veldhuisen, D. J., van Tintelen, J. P., and Jongbloed, J. D. H. (2014). Titin gene mutations are common in families with both peripartum cardiomyopathy and dilated cardiomyopathy. *European heart journal*, 35(32):2165–73.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63.

Vasli, N., Böhm, J., Le Gras, S., Muller, J., Pizot, C., Jost, B., Echaniz-Laguna, A., Laugel, V., Tranchant, C., Bernard, R., Plewniak, F., Vicaire, S., Levy, N., Chelly, J., Mandel, J.-L., Biancalana, V., and Laporte, J. (2012). Next generation sequencing for molecular diagnosis of neuromuscular diseases. *Acta neuropathologica*, 124(2):273–83.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic acids research*, 41(Database issue):D483–9.

Venselaar, H., Beek, T. A. H. T., Kuipers, R. K. P., Hekkelman, M. L., and Vriend, G. (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC bioinformatics*, 11:548.

Vidal, M., Cusick, M. E., and Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6):986–98.

Visscher, P. M. (2008). Sizing up human height variation. *Nature genetics*, 40(5):489–90.

Vitkup, D., Sander, C., and Church, G. M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome biology*, 4(11):R72.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127):1546–58.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164.

Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–8.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, 30(2):159–64.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2016). *gplots: Various R Programming Tools for Plotting Data.* R package version 3.0.1.

Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., and Tieleman, D. P. (2014). Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of chemical theory and computation*, 10(2):676–90.

Webb, B. and Sali, A. (2014). Comparative protein structure modeling using MODELLER. *Current protocols in bioinformatics*, 47:5.6.1–32.

Webb, B. and Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current protocols in protein science*, 86:2.9.1–2.9.37.

Webb, B. A., Forouhar, F., Szu, F.-E., Seetharaman, J., Tong, L., and Barber, D. L. (2015). Structures of human phosphofructokinase-1 and atomic basis of cancer-associated mutations. *Nature*, 523(7558):111–4.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature reviews. Genetics*, 15(11):722–33.

Weinkam, P., Chen, Y. C., Pons, J., and Sali, A. (2013). Impact of mutations on the allosteric conformational equilibrium. *Journal of molecular biology*, 425(3):647–61.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Witham, S., Takano, K., Schwartz, C., and Alexov, E. (2011). A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins*, 79(8):2444–54.

Wójcikowski, M., Ballester, P. J., and Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific reports*, 7:46710.

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–5.

Xue, L. C., Dobbs, D., and Honavar, V. (2011). HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, 12:244.

Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., and Bruford, E. A. (2017). Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research*, 45(D1):D619–D625.

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. E. (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *Journal of molecular biology*, 426(14):2692–701.

Yates, C. M. and Sternberg, M. J. E. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *Journal of molecular biology*, 425(21):3949–63.

Yi, S., Lin, S., Li, Y., Zhao, W., Mills, G. B., and Sahni, N. (2017). Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nature reviews. Genetics*, 18(7):395–410.

Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–10.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature reviews. Genetics*, 16(7):409–20.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics*, 9:40.

Zhang, Z., Miteva, M. A., Wang, L., and Alexov, E. (2012). Analyzing effects of naturally occurring missense mutations. *Computational and mathematical methods in medicine*, 2012:805827.

Zimmermann, M. T., Urrutia, R., Oliver, G. R., Blackburn, P. R., Cousin, M. A., Bozeck, N. J., and Klee, E. W. (2017). Molecular modeling and molecular dynamic simulation of the effects of variants in the TGFBR2 kinase domain as a paradigm for interpretation of variants obtained by next generation sequencing. *PloS one*, 12(2):e0170822.

Appendix A

Supplementary data

Supplementary data can be downloaded at https://github.com/AnnaLaddach/Thesis_Anna_Laddach_2019_Appendix-A.git

This consists of:

1. The impact of using different cut-offs for the minimum number of protein core residues.
2. Details of the number of SAVs which localise to different protein regions in the gnomAD common and rare, COSMIC and Clinvar datasets.
3. A list of pathways annotated by functional cluster ("proliferative", "nucleotide processing" and "response").
4. Proteins enriched in COSMIC non-driver variants at protein-protein interaction sites.
5. Statistics for comparisons of structural network features between datasets.
6. Numbers of proteins and SAVs which underlie correlations between proteomic/transcriptomic features and variant enrichment.
7. Pairwise Spearman correlations between all studied proteomics and transcriptomics features.
8. Enrichment of functional pathways by proteomics/transcriptomic features.
9. Properties of trajectories can be represented by 101 snapshots.